



A General Introduction to the German Socio-Economic Panel Study (SOEP)

- Design, Contents and Data Structure [waves A-V, 1984-2005] -

Contact:

Dr. Joachim R. Frick
SOEP – Data Operations Manager
DIW Berlin (German Institute for Economic Research)
Koenigin-Luise-Str. 5
D-14195 Berlin
Germany
tel.: ++49-(0)30 -89 789-279
fax.: ++49-(0)30 -89 789-109
e-mail: JFRICK@DIW.DE
<http://www.diw.de/gsoep/>

Introduction to the German *Socio-Economic Panel* Study (SOEP)

- Part I : Design, Contents, Sample -

General information

1. Documentation and information material on SOEP and User Support via List Server and E-Mail
2. Supporting software packages
3. Selected references

Contents of the study

4. Aims of the SOEP
5. Main focus and special topical modules since 1984
6. The time dimensions

The Sample of the study

7. “West-German” residents, „Foreigners“, „East-Germans“, „Immigrants“, „Refreshment Sample“, „Innovation Sample“, “High Income Sample“

Survey-related issues

8. Interview methodology
9. Survey instruments
10. The follow-up concept
11. The emergence of “new” households

Development of sample size

12. Starting sample size (Wave 1)
13. Determinants of sample development
14. Development in a cross-sectional perspective
15. Development in a longitudinal perspective
(see SOEPinfo-WWW <http://panel.gsoep.de/soepinfo2004/> section “Basic Information / sample size overview: cross-section/longitudinal”)

General information (1)

Documentation and Information material on the SOEP

- Desktop Companion DTC (on CD-ROM and download via <http://www.diw.de/soep>) with a general introduction to the SOEP containing some exemplary data retrievals for different software packages (TDA, SPSS, SAS, Stata)
- Complete documentation of generated and status variables
- Link to original questionnaires (Englisch and German versions)
- SOEP-NEWSLETTER published quarterly in a paper version; also available via SOEP-Homepage <http://www.diw.de/gsoep>

User support via List Servers and hotlines for German and International Users

International Users

- Electronic listserver at Cornell University

In order to join the Cornell English language listserver for users of the SOEP and CNEF data, please send an email to the following address

listproc@cornell.edu

and in the ****Body**** of the email, include the following **one** line:

subscribe GSOEP-L gsoep

We suggest erasing any "signature" you might have at the bottom of your email.

- E-mail hotline at Cornell University GSOEP@cornell.edu.

German Users

- Electronic listserver at DIW Berlin

If you want to join the SOEP listserver at DIW Berlin, please send an email to:

sympa@list.diw.de

Subject: subscribe soep-l:

For security reasons you will receive an automated email asking for confirmation.

- E-mail hotline at DIW Berlin soepmail@diw.de.

General information (2)

Supporting Software Packages

- *SOEPinfo*: web-based user support system accessible via SOEP-Homepage <<http://www.diw.de/gsoep>> including
 - an item-correspondence list of all available SOEP-variables with frequencies (information includes variable and value labels, names, and the file in which they are stored).
 - a link to the corresponding questions in the original survey instrument (in German and English language)
 - a tool for generating program syntax files for retrieving SOEP micro-data (available for major statistical software packages SPSS, Stata, SAS)

- *SOEPLIT* (via SOEP-homepage and CD-ROM) contains a list of published and unpublished references of analyses using SOEP-data (as of May 2006: approx. 3,700 entries).

- **INSTALLATION TOOL** on CD-ROM supports users when installing SOEP data on their hard disk. It allows one to select popular software formats (SPSS, SAS, Stata, ASCII) as well as the number of files/waves to install. However, we strongly suggest to always install **all** files of any new data release.

- **NEWSPELL** is a program to rearrange the original survey data stored in spell-files like PBIOSPE, ARTKALEN, EINKALEN, BIOMARSY, etc.. It allows users to define (a) a new priority of spelltypes by recoding and aggregating given SPELLTYP information and (b) the period of time which is of interest for a given analysis. Optional output is (a) spell-data according to self-defined priority and time period including information on previous and next spelltype and (b) time-series data according to self-defined priority and time period.

- **BIOSCOPE** allows graphical viewing of spell information for single persons. This data includes the main activity information in the file PBIOSPE and covers events and duration of training, employment, unemployment, retirement etc.

General information (3)

Selected References

General Overview (SOEP, CNEF)

- Burkhauser, Richard V, Barbara A. Butrica, Mary C. Daly and Dean R. Lillard (2001): The Cross-National Equivalent File: A product of cross-national research. In: Becker, Irene, Ott, Notburga and Rolf, Gabriele (Eds.) Soziale Sicherung in einer dynamischen Gesellschaft (Social Insurance in a Dynamic Society). Festschrift für Richard Hauser zum 65. Geburtstag (Papers in Honor of the 65th Birthday of Richard Hauser), Campus, Frankfurt/New York
- Haisken-DeNew, John P. and Frick, Joachim R. (2005): Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 8.0 – Update to Wave 21. German Institute for Economic Research, Berlin.
- SOEP-Group (2001): The German Socio-Economic Panel Study (GSOEP) after more than 15 years – Overview. In: Holst, Elke, Dean R. Lillard and Thomas A. DiPrete (Eds.): Proceedings of the 2000 Fourth International Conference of Socio-Economic Panel Study Users. Vierteljahrshefte zur Wirtschaftsforschung, 70 (1), 134-144.
- Wagner, Gert; Burkhauser, Richard V. and Behringer, Friederike (1993): The English Language Public Use File of the German Socio-Economic Panel Study. In: The Journal of Human Resources, 28(2), 429-433.

Selected Research (incl. Proceedings of SOEP conferences)

- Büchel, Felix†, Conchita D'Ambrosio, and Joachim R. Frick (eds.), Proceedings of the "6th International Conference of German Socio-Economic Panel Study Users (SOEP2004). Schmollers Jahrbuch - Zeitschrift für Wirtschafts- und Sozialwissenschaften 125 (1), 2005
- Burkhauser, Richard V. and Gert G. Wagner (eds.), Proceedings of the 1993 International Conference of German Socio-Economic Panel Study Users. Vierteljahrshefte zur Wirtschaftsforschung 63 (1/2), Berlin: Duncker & Humblot, 1994.
- Dunn Thomas A. and Johannes Schwarze (eds.), Proceedings of the 1996 Second International Conference of the German Socio-Economic Panel Study Users. Vierteljahrshefte zur Wirtschaftsforschung, 66(1), Berlin: Duncker & Humblot, 1997.
- Dunn, Thomas A., Joachim R. Frick and James C. Witte (eds.), Proceedings of the 1998 Third International Conference of the German Socio-Economic Panel Study Users. Vierteljahrshefte zur Wirtschaftsforschung, 68(2), Berlin: Duncker & Humblot, 1999.
- Frick, Joachim R. and Grabka, Markus M. (2003): Imputed Rent and Income Inequality: A Decomposition Analysis for the U.K., West Germany, and the USA. *Review of Income and Wealth*. 49(4): 513-537.
- Goodin, Robert E., Bruce Headey, Ruud Muffels, and Henk-Jan Dirven, The Real Worlds of Welfare Capitalism. Cambridge University Press, 1999.
- Holst, Elke, Dean R. Lillard and Thomas A. DiPrete (eds.), Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP 2000). Vierteljahrshefte zur Wirtschaftsforschung, 70(1), Berlin: Duncker & Humblot, 2001.
- Holst, Elke, Jennifer Hunt and Jürgen Schupp (eds.), Proceedings of the 5th International Conference of Socio-Economic Panel Users. Schmollers Jahrbuch, 123 (1), Berlin: Duncker & Humblot, 2003.
- Krause, Peter, Gerhard Bäcker, and Walter Hanesch, Combating Poverty in Europe: The German Welfare Regime in Practice (Studies in cash and care). Aldershot: Ashgate, 2003
- Schwarze, Johannes, Friedrich Buttler and Gert Wagner (eds.), Labor Market Dynamics in Present Day Germany. Frankfurt/Main, New York: Campus; Boulder, Colorado: Westview Press, 1994.
- Schyns, Peggy: Income and life satisfaction - A cross-national and longitudinal study (PhD thesis). Delft: Eburon, 2003.
- Van Praag, Bernard M.S. and Ada Ferrer-i-Carbonell, Happiness Quantified - A Satisfaction Calculus Approach. Oxford: Oxford University Press, 2004.

Samples, Weighting and Imputation

See SOEP-website for documentation:

<http://www.diw.de/english/sop/service/doku/index.html#1.2>

Frick, Joachim R. and Grabka, Markus M. (2004): Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution. DIW-Discussions Papers No. 376-revised, April 2004, Berlin: DIW Berlin.

Frick, Joachim R. and Grabka, Markus M. (2005): Item-non-response on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv*, 89: 49-60

Infratest Sozialforschung (1984-2004): Das Sozio-ökonomische Panel, Methodenbericht zur Haupterhebung, München.

Kroh, Martin and Spiess, Martin (2005): Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) 1984 – 2004, DIW Berlin Data Documentation No. 6, Berlin: DIW Berlin.

http://www.diw.de/deutsch/produkte/publikationen/datadoc/docs/diw_datadoc_2005-006.pdf

Pannenberg, Markus (2002): Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (GSOEP) (1984 until 2001). Research Notes Nr. 23, Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).

Pischner, Rainer (1994): Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels. In: Gabler, Siegfried; Hoffmeyer-Zlotnik, Jürgen H.P. und Krebs, Dagmar (Hrsg.), Gewichtung in der Umfragepraxis, Opladen: Westdt. Verlag, S. 166-187.

Rendtel, Ulrich (1995): Lebenslagen im Wandel: Panalausfälle und Panelrepräsentativität, Frankfurt/M. - New York: Campus.

Rendtel, Ulrich, Markus Pannenberg und Stefan Daschke (1997) Die Gewichtung der Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP), in: Vierteljahrshefte zur Wirtschaftsforschung, Heft 2, Vol. 66, S.271-285.

Rendtel, Ulrich; Wagner, Gert und Frick, Joachim (1995): Eine Strategie zur Kontrolle von Längsschnittgewichtungen in Panelerhebungen - Das Beispiel des Sozio-Oekonomischen Panels (SOEP). In: *Allgemeines Statistisches Archiv*, Jg. 79, Heft 3, S. 252-277.

Documentation of Generated Variables

See SOEP-website for complete documentation:

<http://www.diw.de/english/sop/service/doku/index.html>

Frick, Joachim R. and Thorsten Schneider (2004): Biography and Life History Data in the German Socio Economic Panel (SOEP), DIW Berlin.

<http://www.diw.de/english/sop/service/doku/index.html#1.3>

Contents of the Study (1)

- *Field Title:* “Living in Germany”
- *Aims of the SOEP:*
 - collect yearly representative microdata on persons, households and families in the Federal Republic of Germany since 1984 (in East Germany since 1990) ...
 - to measure stability and change in living conditions ...
 - by following principally a micro-economic approach ...
 - enriched with sociological and political science variables, mainly determined by the “Social Indicator movement”:
 - objective indicators (e.g. income, employment status)
 - subjective indicators to measure the individual perception of objective living conditions (e.g. satisfaction, values, preferences).
- *Standard Components (measured (bi-) yearly)*
 - Demography and Population
 - Labor Market and Occupation
 - Income, Taxes, and Social Security
 - Housing
 - Health
 - Household Production
 - Education, Training, and Qualification
 - Basis Orientation (preferences and values), Participation, and Integration

Contents of the Study (2)

- *Special topical modules*

Data in distribution

Year	Wave	Sample	Topic	
1984	A	1	A,B	Employment biography since age 15 (Bio)
1985	B	2	A,B	Marriage and family biography (Bio)
1986	C	3	A,B	Social origins (Bio), first job (Bio), neighborhood
1987	D	4	A,B	Social security, early retirement, persons requiring care, and child care
1988	E	5	A,B	Assets
1989	F	6	A,B	Further education or training and qualification
1990	G	7	A,B	Use of time and preferences
1991	H	1	C	Base questions (labor market, subjective indicators)
		8	A,B	Family and social services
		2	C	Family and social services (shortened version plus relication of subjective indicators and labor market indicators of wave 1 base questions)
		9	A,B	Social security and poverty (partly replication Wave D)
1992	I	3	C	Social security and poverty, labor market indicators and biographical information (Bio)
		10	A,B	Further education or training (short. replication wave F)
1993	J	4	C	Further education or training, labor market
		11/5	A,B,C	Neighborhood, values, and expectations
1994	K	1	D1	Same plus immigration history and biography
		12/6	A,B,C,D1	Partial replication of Wave G - use of time and preferences, increased range of income questions
1995	L	1	D2	Same plus immigration history and biography
		13/7/2	A,B,C,D	Replication of social network questions (Wave H)
1996	M	14/8/3	A,B,C,D	Social security and poverty (replication of Wave I)
1997	N	15/9/4/1	A,B,C,D,E	Ecology and environmental behavior (indirect taxation)
1998	O	16/10/5/2	A,B,C,D,E	Neighborhood, Values, Expectations
1999	P	17/11/6/3/1	A,B,C,D, E,F	Further education, Training, Labor Market
2000	Q	18/12/7/4/2	A,B,C,D, E,F	Social Network, Working conditions (see also Waves H, M)
2001	R	19/13/8/5/3	A,B,C,D, E,F, G	Wealth (see also Wave E), Replication Social security (Wave N)
2002	S	20/14/9/6/4	A,B,C,D, E,F,G	Ecology and Environmental Behavior (see also Wave O); Bio-Data for Sample G
2003	T	21/15/10/7/5	A,B,C,D, E,F,G	Further Education or Training; Qualification (see also Waves F, J, Q)
2004	U	22/16/11/8/6	A,B,C,D, E,F,G	Use of Time and Preferences (see also Waves G, L)
2005	V	23/17/12/9/7	A,B,C,D, E,F,G	Family and social services, networks (see also Wave H, M, R)
2006	W		A,B,C,D, E,F,G	

	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	
<i>Employment biography</i>	A+B	R	R	R	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO
<i>Marriage and family biography</i>		A+B	R	R	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO
<i>Social origin, first job</i>			A+B	R	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO	BIO
<i>Neighborhood</i>			A+B								A+B /C					A+B /C/D /E							
<i>Social security and its perception</i>				A+B					A+B /C					A+B /C/D					A+B /C/D /E/F/ G				
<i>Assets</i>					A+B																		
<i>Further education or training and qualification</i>						A+B				A+B /C							A+B /C/D /E					A+B /C/D /E/F/ G	
<i>Use of time and preferences</i>							A+B					A+B /C/D					A+B /C/D /E/F						A+B /C/D /E/F/ G
<i>Family and social services</i>								A+B /C					A+B /C/D						A+B /C/D /E/F				
<i>Working (job) conditions</i>		A+B	A+B	A+B								A+B /C/D							A+B /C/D /E/F				
<i>Future Expectations</i>											A+B /C					A+B /C/D /E							
<i>Ecology and environmental behaviour</i>															A+B /C/D						A+B /C/D /E/F/ G		

Legend:

A = Sample A; B = Sample B; C = Sample C; D = Sample D; E = Sample E; F=Sample F; G=Sample G

BIO = Biography questionnaire to be answered by first time respondents

R = Retrospectively asked biographical information

Contents of the Study (3)

In order to measure change in a longitudinal study one has to consider different *dimensions of time*: Thus the SOEP contains questions measuring time in several ways.

- Questions about a point of time (present time)
e.g. current employment status or current levels of satisfaction
- Single retrospective questions on certain events in the past (past time)
e.g. how often did you change your job during the last ten years?
- Retrospective life event history since the age of 15 (past time)
e.g. employment or marital history
- Monthly calendar on income and labor market related issues (past time)
e.g. employment status January through December last year
- Questions concerning a period of time (past time)
e.g. demographic changes since the last interview like marriage or death of spouse
- Questions concerning future prospects (future)
e.g. satisfaction with life five years from now, or job expectations

The SOEP Sample

The SOEP (as of 2002) contains data on seven different subsamples. Each of these was drawn in a different multi-step random sampling process.

- A** “West-German” residents: started in 1984
 - n=4528 or 4298 households*
 - Head of household is either German or of another nationality than those in Sample B.

- B** “Foreigners”: started in 1984
 - n=1393 or 1326 households*
 - Head of household is either Turkish, Italian, Spanish, Greek, or Yugoslavian.

- C** “East-Germans”: started in 1990
 - n=2179 or 2071 households*
 - Head of household at the time of the survey was a citizen of the GDR.

- D** “Immigrants”: started in 1994/95 in two different subsamples
 - 1994: Subsample D1 with n=236 households
 - 1995: Subsample D2 with n=295 households
 - total in 1995 n=522 or 497 households* (D1 and D2)
 - At least one household member has moved from abroad to Germany after 1984.

- E** “Refreshment sample”: started in 1998
 - n=1067 or 1014 households*
 - Random sample covering all existing subsamples (total population)

- F** “Innovation sample”: started in 2000
 - n=6052 or 5750* households
 - Random sample covering all existing subsamples (total population)

- G** “High Income Sample”: started in 2002
 - n=1224 or 1163* households
 - Household monthly net income > 7.500 DM (threshold has been revised in wave 2 to 4.500 EURO yielding a decrease in households to be followed up)

*The first number relates to the full 100% version, the second relates to the 95% scientific use version of the SOEP-data (often referred to GSOEP).

SOEP Subsamples [observation years and waves]

		'84	'85	'86	'87	'88	'89	'90	'91	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05
A	“West-Germans”	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
B	“Foreigners”	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
C	“East-Germans”							1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
D1	“Immigrants”											1	2	3	4	5	6	7	8	9	10	11	12
D2	“Immigrants”												1	2	3	4	5	6	7	8	9	10	11
E1	Refreshment („PAPI“)															1	2	3	4	5	6	7	8
E2	Refreshment („CAPI“)															1	2	3	4	5	6	7	8
F	„Innovation“																	1	3	3	4	5	6
G	“High Income”																			1	2	3	4

Survey-related issues (1)

Interview methodology

- In principle, face-to-face individual interviews with all household-members aged 16 and over.
- In principle, paper-and-pencil interviews (subsamples A through D, E1).
- In principle, there are no proxy interviews (third persons being interviewed instead of the respondent him/herself).
- Household interviews with “head of household”. This is the person who knows best about the general conditions under which the household acts. In order to reduce longitudinal inconsistencies this person is supposed to answer the household-questionnaire each year, unless he/she leaves the household or dies.
- An increasing number of individuals completes the questionnaire without the presence of the interviewer. In case of major inconsistencies or missing information the field work agency “TNS Infratest Sozialforschung, Munich” will give the respondent a phone call to resolve the issue.
- Starting with 50% of the refreshment sample E in 1998, CAPI (computer assisted personal interview) was introduced to the SOEP for the very first time. The remaining 50% of sample E interviews are conducted the “old-fashioned way” as paper-and-pencil interviews (PAPI) in order to allow for analyses of mode effects.

====> This “methodology mix” is documented in a set of variables describing the interview situation on the individual level (can be found in the \$P and the \$PGEN files).

Survey-related issues (2)

Survey Instruments:

- **Address log** containing general information (filled in by interviewer; see files \$PBRUTTO for individuals and \$HBRUTTO for households, respectively)
 - on households (e.g. size, housing area, regional information) and individuals therein (e.g. sex, year of birth, relationship to head)
 - on the process of field work (e.g. number of contacts, reason for drop-outs, interview method)
 - different versions according to survey status
 - *old* households at the old address (“green” version)
 - *new* or moved households (“blue” version)
- **Questionnaire** versions with pre-tested questions
 - individual respondents* (see files \$P) vs. households** (see files \$H)
 - different subsamples: “West-Germans” (sample A), “Foreigners” (B), “East-Germans” (C), “Immigrants” (D)
 - survey status: old vs. new or moved households (“green” / “blue” version)
 - questionnaire for temporary drop-outs (see files \$PLUECKE)
 - Age-specific questionnaires
 - „Life History/biography“ (marriage and family, employment, social origin, immigration, etc.)***
 - „youth “ (16-17 years old first time respondents receive this instrument on adolescence instead of the standard biography questionnaire, since 2001)
 - „mother & child “ (asked from mothers of new borns aged 0-15 months, since 2003)
 - „infant“ (asked from mothers of 2-3 year old children, since 2005)

* The individual questionnaire is integrated for all subsamples since 1994. First time respondents (former “blue” version) answer additionally a special questionnaire on biographical issues (employment history, marital history, social origin, etc.).

** The household questionnaire is integrated for new and old households since 1991.

*** The „Life History/biography“ questionnaire is integrated for all samples since 1996.

Survey-related issues (3)

The follow-up concept

- All persons taking part in wave 1 of the survey (respondents as well as their children) are to be surveyed also the next year: at the same address as well as after a move within Germany (covering residential mobility).
- Since all persons are to be personally interviewed once they reach the age of 16, the next generation is automatically taken into account (demographic development).
- Persons moving into an existing SOEP household are to be surveyed (regional mobility). Since 1989 these persons are also followed in case of leaving the household. This had not been the case up to wave 5 (1988).
- Temporary drop-outs: Persons and households which could not be successfully interviewed in a given year are followed until there are two consecutive temporary drop-outs of all household members or a final refusal. In case of a successful interview after a drop-out there is also a small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status).

Survey-related issues (4)

The emergence of new households

		<i>Households</i>	
		Old	New
<i>Persons</i>	Old	<ul style="list-style-type: none">▪ “classic case” without change of address▪ entire household moves	<ul style="list-style-type: none">▪ Move-out
	New	<ul style="list-style-type: none">▪ Birth▪ Move-in	<ul style="list-style-type: none">▪ Birth▪ caused by splitt-offs of old persons from old households*

* Remember that households *new* to the SOEP may already have existed before contacting the survey.

Development of Sample Size (1)

Starting Sample Size in Wave 1

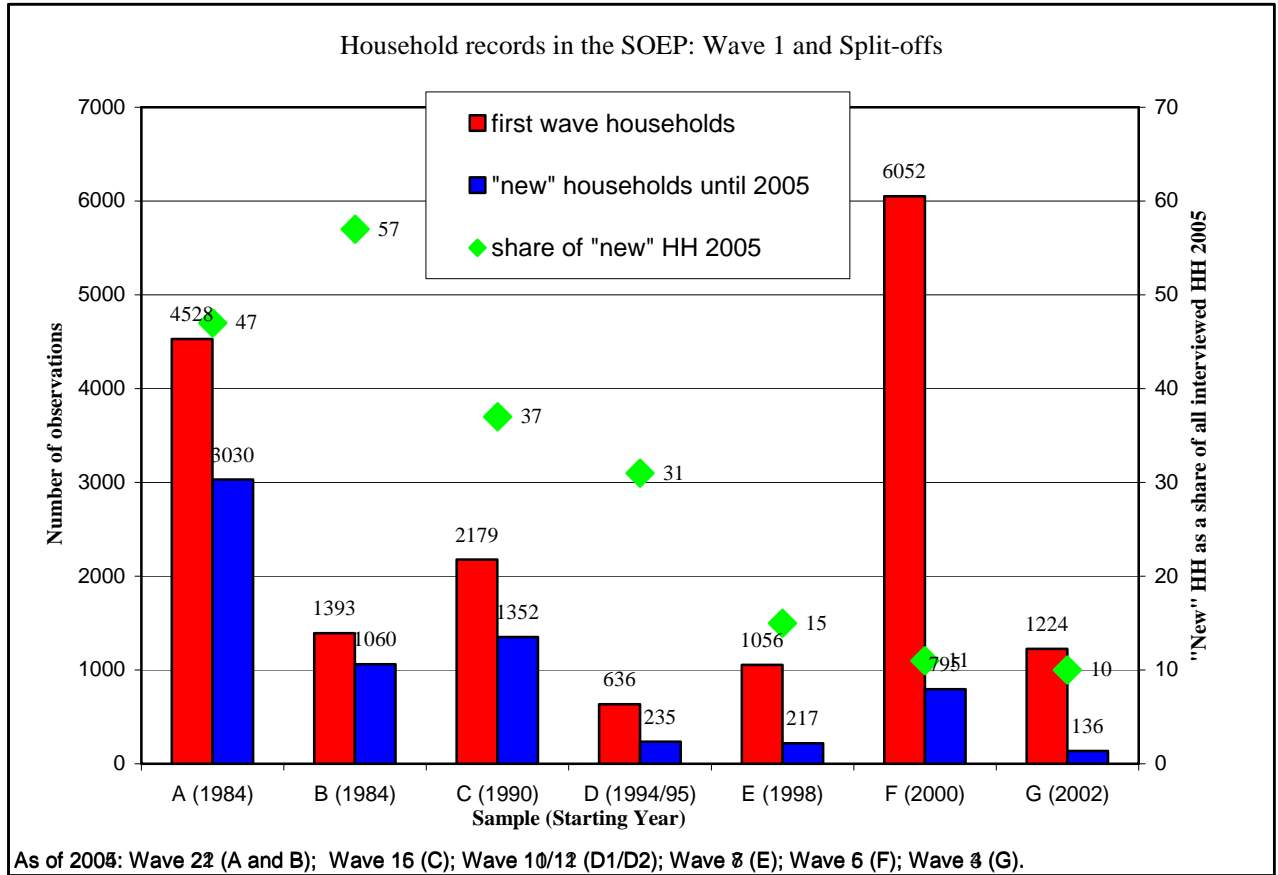
Sample	Year	Households (net)	Persons (gross)	Respondents (net)	Children (net)
100%-Sample					
A and B	1984	5921	16205	12245	3915
C	1990	2179	6131	4453	1591
D1	1994	236	733	471	248
D1/D2	1995	522	1665	1078	517
E	1998	1067	2470	1923	468
F	2000	6052	14525	10890	2993
G	2002	1224	3538	2671	693
95%-Sample					
A and B	1984	5624	15397	11610	3711
C	1990	2071	5818	4229	1510
D1	1994	225	696	451	231
D1/D2	1995	497	1584	1027	488
E	1998	1014	2342	1827	448
F	2000	5750	13772	10324	2838
G	2002	1163	3359	2536	653

Development of Sample Size (2)

Determinants of the Sample Development

- (1) Demographic factors
 - Persons exit by:
 - Death
 - Moving abroad
 - Persons enter by:
 - Birth
 - Moving into a SOEP household from somewhere else in Germany or from abroad
 - Reaching age of 16 years (minimum respondents age is given by the calendar year, in which a person turns 17 years of age)
 - Split-offs of at least one *old* person from an *old* household
- (2) Field-work related factors (2 stages)
 - making a successful contact to a given household
 - realizing a successful interview
 - social groups typically associated with problems in respect to re-contacting and re-interviewing:
 - single person households;
 - mobile households and persons;
 - young adults leaving parental home
- (3) Panel maintenance
 - For each successful interview, any respondent
 - receives a gift related to the yearly topical module
 - takes part in a monthly nationwide lottery.
 - Addresses are kept up to date by the field work agency throughout the entire year in order to be informed about residential mobility; for example by sending them a brochure containing some results of analyses based on previous SOEP data.
 - The interview situation (face-to-face) ensures a personal relationship, which makes it harder to withdraw from the survey. Thus, the stability of the interviewer over time is very crucial.

▪ **“Old” vs. “new” households in the SOEP**

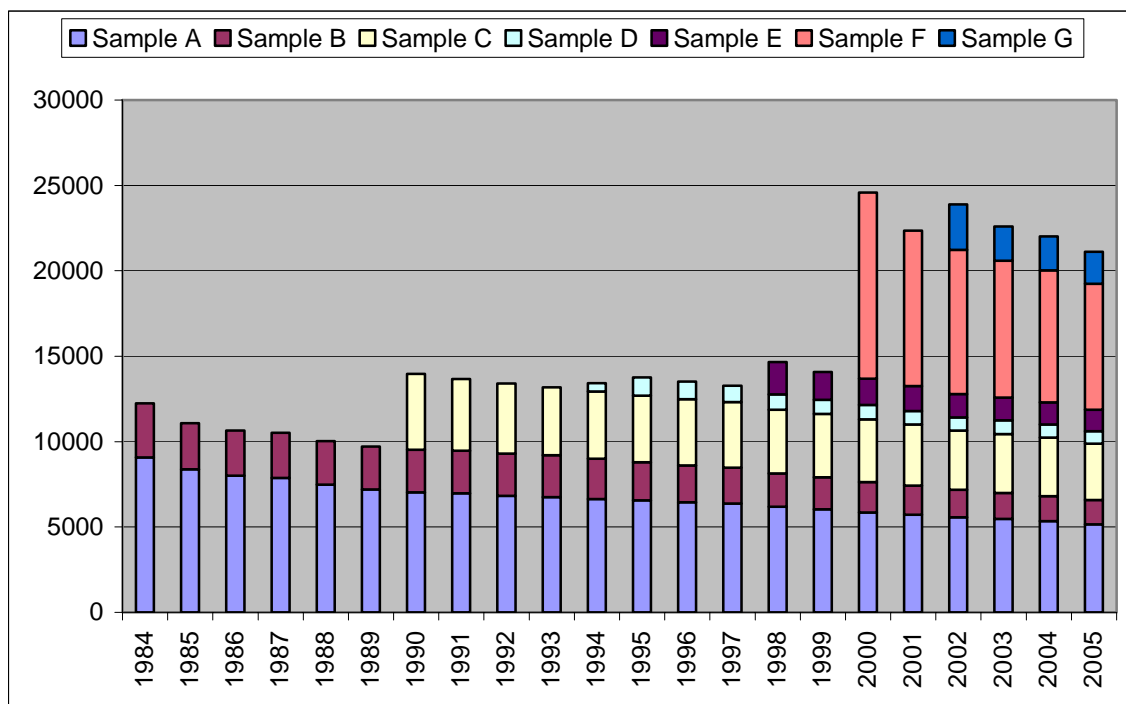


Development of Sample Size (3)

Cross-sectional perspective (based on 100% sample):

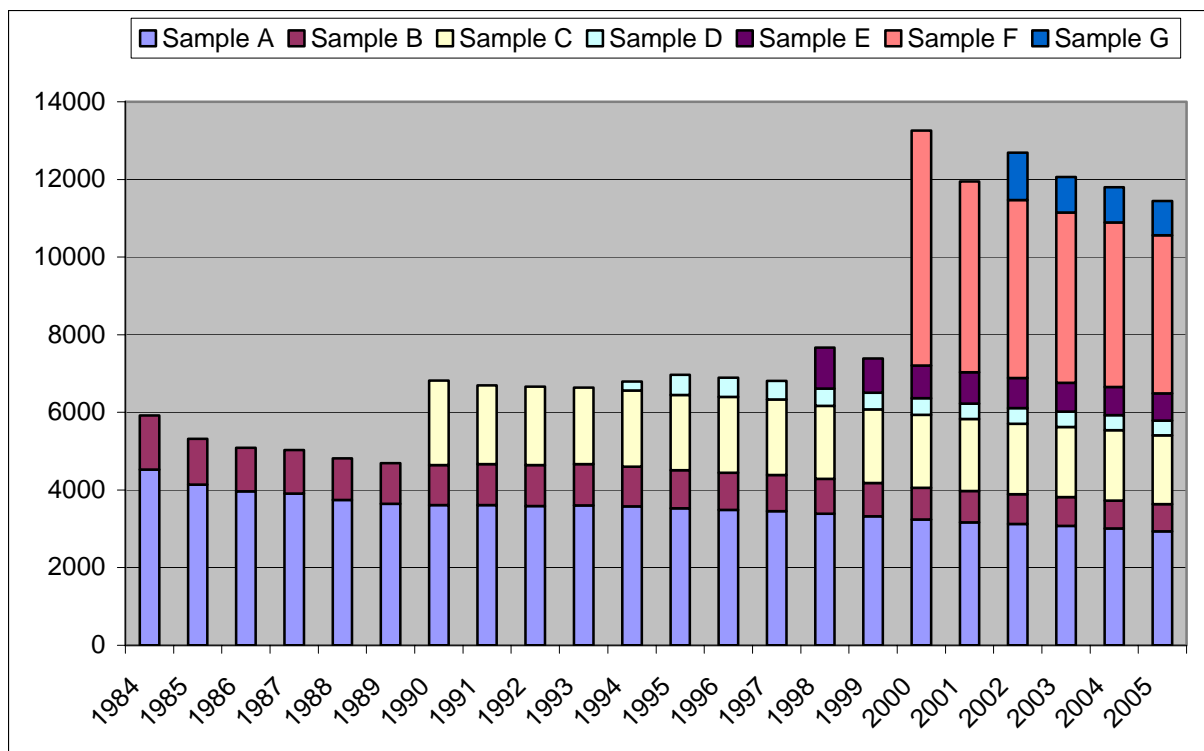
Number of successfully interviewed persons by sample (files \$P)

Year	Sample							Total
	A	B	C	D	E	F	G	
1984	9076	3169						12245
1985	8372	2718						11090
1986	8009	2637						10646
1987	7868	2648						10516
1988	7481	2542						10023
1989	7201	2509						9710
1990	7036	2483	4453					13972
1991	6974	2493	4202					13669
1992	6821	2484	4092					13397
1993	6747	2459	3973					13179
1994	6637	2364	3945	471				13417
1995	6567	2231	3892	1078				13768
1996	6454	2152	3882	1023				13511
1997	6378	2089	3844	972				13283
1998	6184	1961	3730	885	1910			14670
1999	6045	1864	3709	838	1629			14085
2000	5852	1771	3687	837	1549	10890		24586
2001	5713	1711	3576	789	1464	9098		22351
2002	5577	1598	3466	780	1373	8427	2671	23892
2003	5480	1519	3453	789	1332	8006	2013	22592
2004	5343	1468	3435	760	1300	7727	1986	22019
2005	5152	1423	3311	735	1241	7372	1871	21105



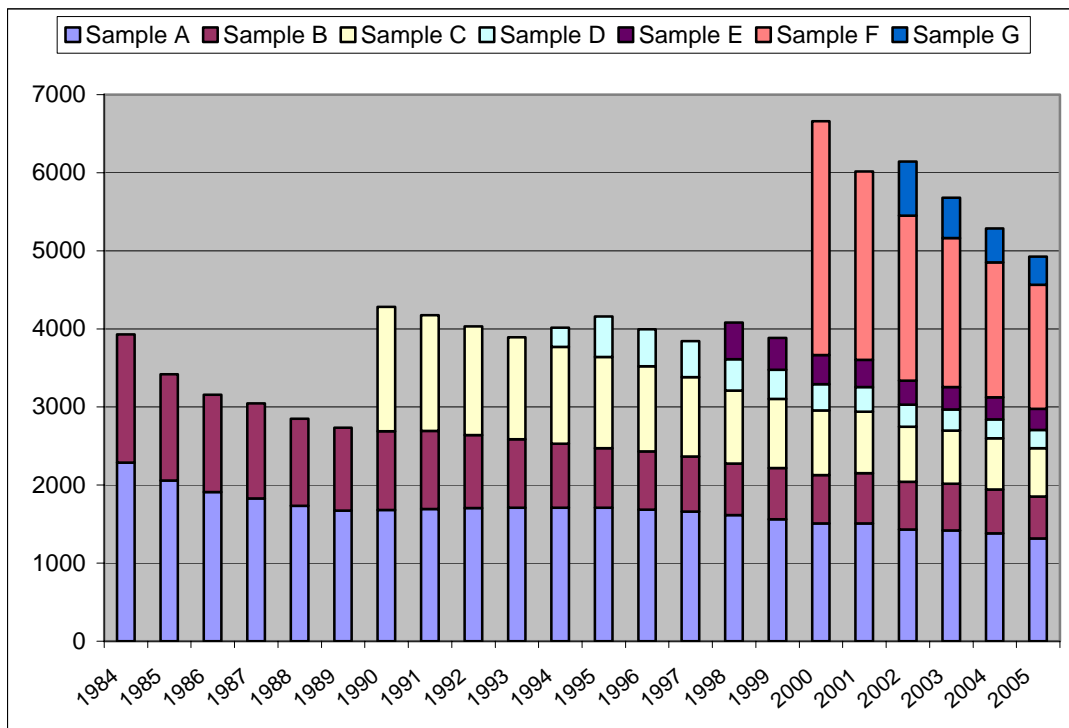
Number of successfully interviewed households by sample (files \$H)

Year	Sample							Total
	A	B	C	D	E	F	G	
1984	4528	1393						5921
1985	4141	1181						5322
1986	3962	1128						5090
1987	3910	1116						5026
1988	3743	1071						4814
1989	3647	1043						4690
1990	3612	1028	2179					6819
1991	3613	1056	2030					6699
1992	3585	1060	2020					6665
1993	3603	1064	1970					6637
1994	3577	1023	1959	236				6795
1995	3526	982	1938	522				6968
1996	3485	960	1951	498				6894
1997	3458	931	1942	479				6810
1998	3387	898	1886	441	1056			7668
1999	3325	858	1894	425	886			7388
2000	3240	820	1879	425	842	6052		13258
2001	3168	809	1850	398	811	4911		11947
2002	3123	766	1818	402	773	4586	1224	12692
2003	3072	742	1807	399	744	4386	911	12061
2004	3010	714	1813	388	732	4235	904	11796
2005	2937	698	1771	379	706	4070	879	11440



Number of children (up to 16 years of age)
in successfully interviewed households by sample (files \$KIND)

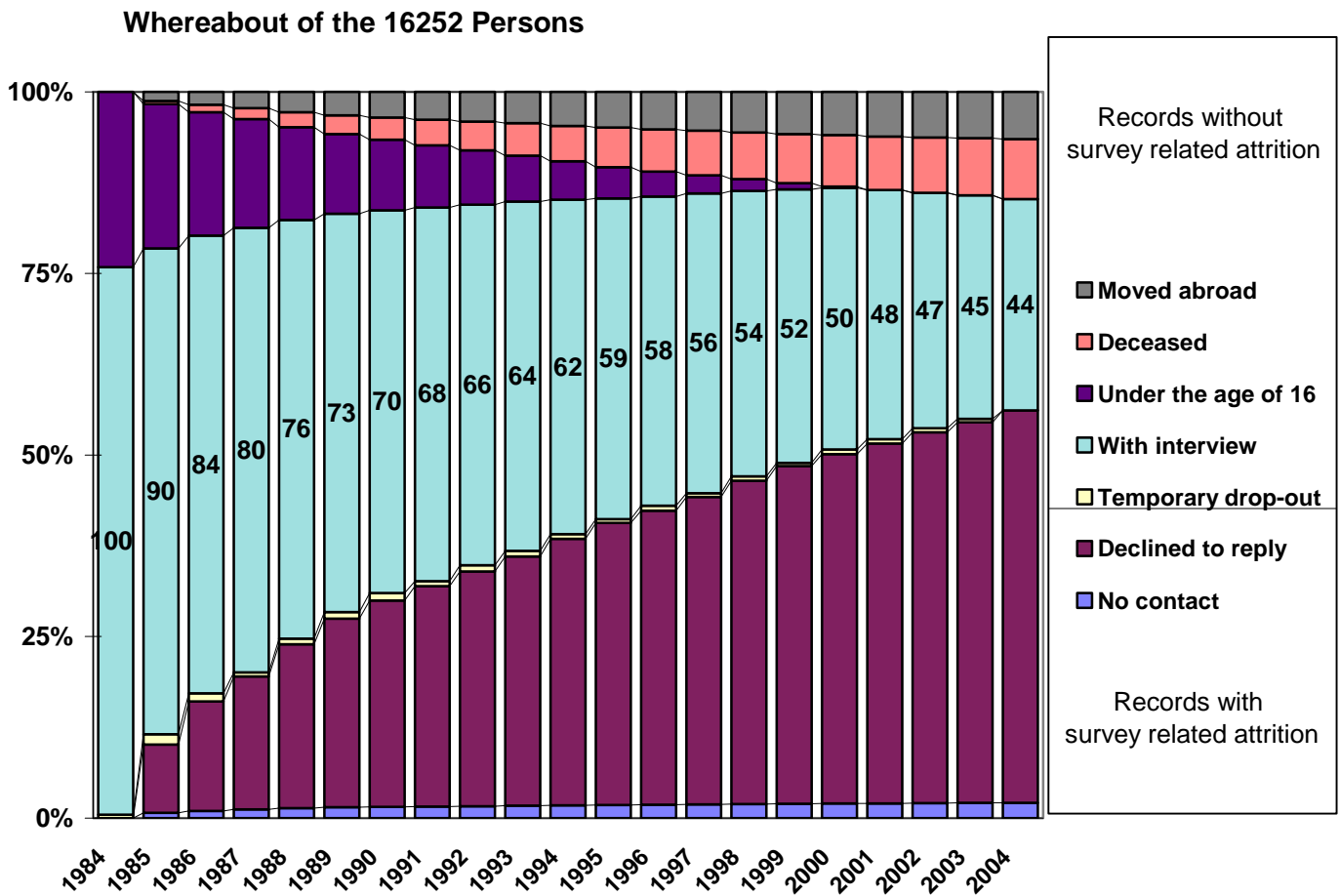
Year	Sample							Total
	A	B	C	D	E	F	G	
1984	2290	1638						3928
1985	2058	1360						3418
1986	1913	1245						3158
1987	1828	1219						3047
1988	1736	1113						2849
1989	1675	1059						2734
1990	1681	1010	1591					4282
1991	1693	1001	1481					4175
1992	1707	932	1393					4032
1993	1710	879	1303					3892
1994	1709	823	1236	248				4016
1995	1709	763	1169	517				4158
1996	1684	746	1094	469				3993
1997	1661	707	1017	458				3843
1998	1615	659	937	403	466			4080
1999	1564	656	884	375	406			3885
2000	1508	621	826	339	372	2993		6659
2001	1508	644	790	315	348	2412		6017
2002	1431	612	706	280	308	2111	693	6141
2003	1419	598	681	273	283	1908	516	5678
2004	1381	563	655	243	283	1726	434	5285
2005	1316	536	620	235	271	1586	359	4923



Development of Sample Size (4)

Longitudinal perspective 1984-2005 (Waves A – V):

Samples A and B (SOEP population as of 1984)



Introduction to the German *Socio-Economic Panel* Study (SOEP)

- Part II: The Data -

Data structure

1. Cross-sectional (yearly) datasets
2. Longitudinal datasets
3. How do cross-sectional and longitudinal datasets relate to each other ?

Linking data across time

4. The set of identifiers: households, individuals, spells
5. Identifiers in the course of time

Dealing with Variables

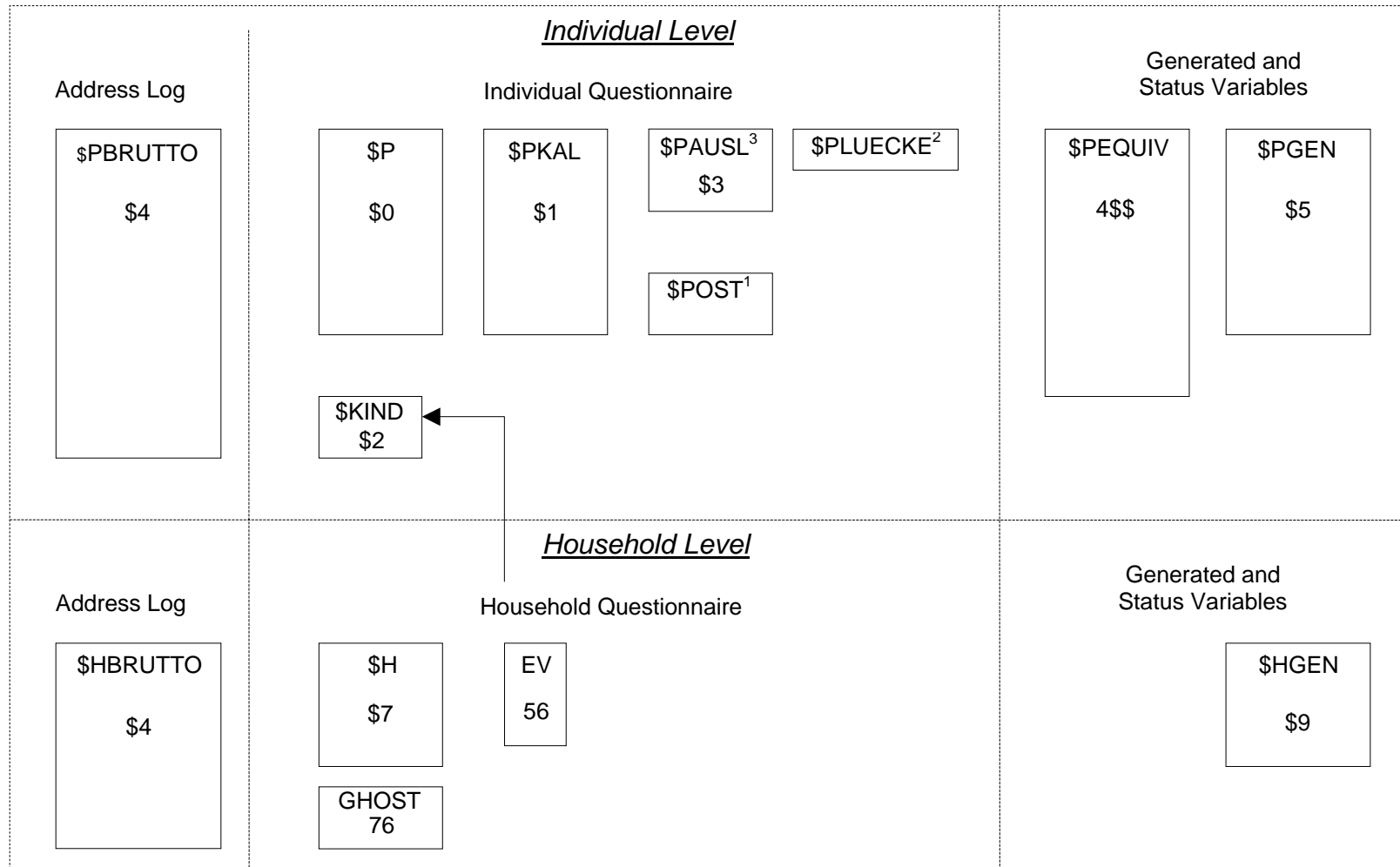
6. Principles for naming survey variables
7. Missing values

Preparing data for analysis (Optional data-structures)

8. Cross-sectional structure
 - a) single cross-sectional data
 - b) comparison of different cross-sections
 - c) pooling cross-section data
9. Longitudinal structure
 - a) complete case analysis (balanced panel design)
 - b) down- and upstream models
 - c) complete information analysis with incomplete data (unbalanced panel design)
 - d) pooled longitudinal data
10. Event or spell data

The SOEP data structure (1)

- Cross-sectional datasets -



\$: Wave specification: A, B, C... V for file names; 1, 2, 3 ... 22 for file numbers.

¹ Waves G and H only; ² Waves B through Q only; ³ Waves A through L only

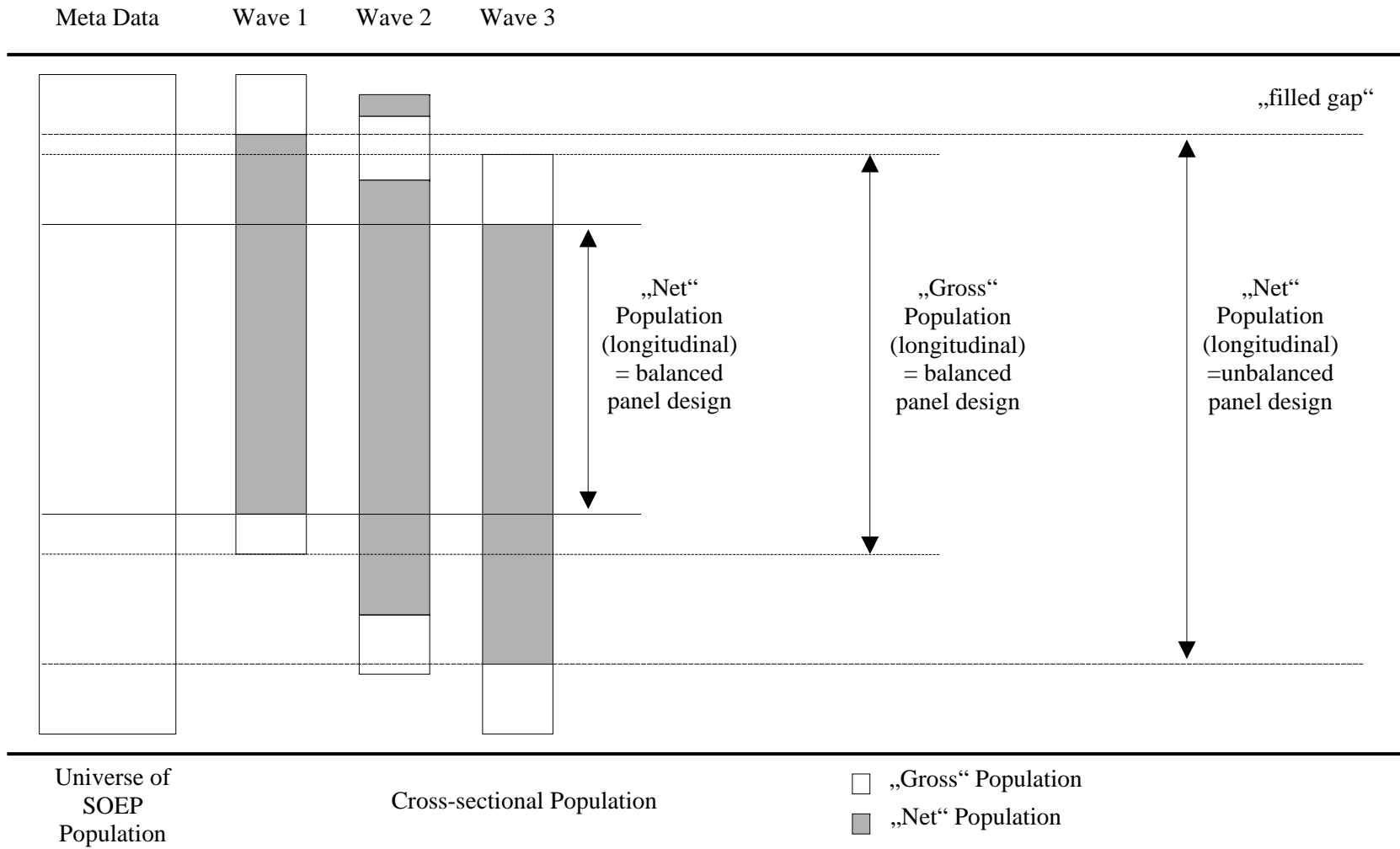
The SOEP data structure (2)

- Longitudinal datasets -

<i>Individual/ Household</i>				<i>Cumulative Individual Data</i>	<i>Spell</i>					<i>Individual</i>			
META-DATA		WEIGHTING FACTORS		DROP-OUTS	SOCIAL ASSIST. (month)	CALENDAR (month)		OCCUP. BIO (year)	MARITAL STATUS (year) (month)		BIRTH (women & men)	PARENTAL INFO	
PPFAD 01	HPFAD 02	PHRF 03	HHRF 07	YPBRUTTO 260	SOZ-KALEN 289	ART-KALEN 81	EIN-KALEN 82	PBIO-SPE 80	BIO-MARSY 282	BIO-MARSM 281	BIO-BIRTH 280	BIO-PAREN 283	
				PERSONS NEEDING CARE							BIO-BRTHM 291		
											BIORESID 288	2nd RESIDENCE	
											AGE SPEC. INFO		
											BIOSOC 287	BIO-AGE17 297	BIOAGE01 293
												BIOAGE03 294	
				PFLEGE 268							BIOTWIN 292	MULTIPLES	
											MIGRATION (migrants only)	FIRST JOB	
											BIO-IMMIG 284	BIO-JOB 285	
SAMP 909		VARIANZ 450											

The SOEP data structure (3)

- Cross-Sectional and Longitudinal datasets -



Linking data across time (1)

- The set of identifiers: households, individuals, spells

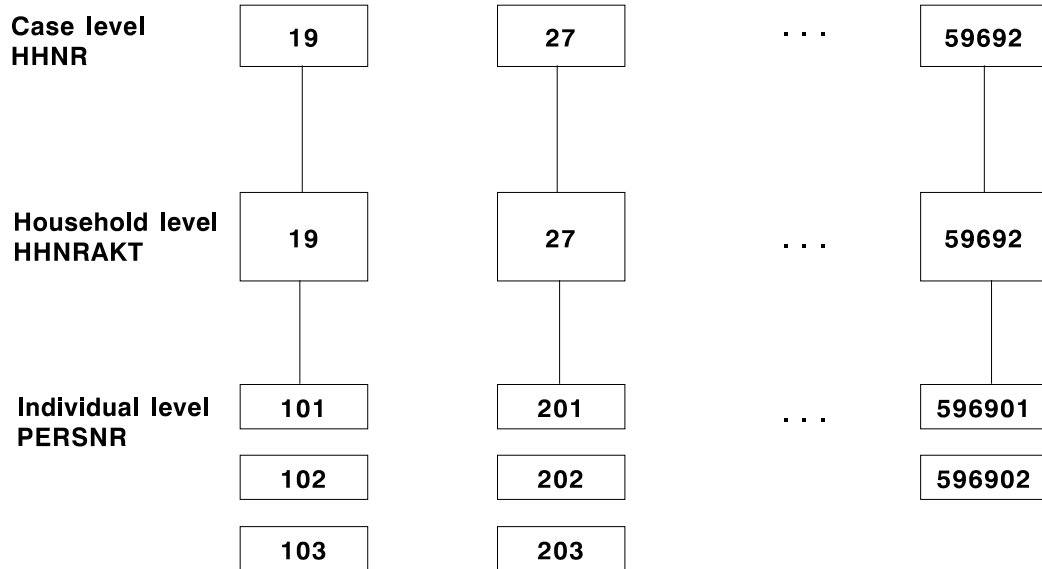
Filename	Case-ID	Sort-ID 1	Sort-ID 2
PPFAD	HHNR	PERSNR	-
PHRF	HHNR	PERSNR	-
HPFAD	HHNR	HHNRAKT	-
HHRF	HHNR	HHNRAKT	-
\$PBRUTTO	HHNR	HHNRAKT*	PERSNR
\$P	HHNR	HHNRAKT*	PERSNR
\$PAUSL	HHNR	HHNRAKT*	PERSNR
\$PGEN	HHNR	HHNRAKT*	PERSNR
\$PKAL	HHNR	HHNRAKT*	PERSNR
\$PLUECKE	HHNR	HHNRAKT*	PERSNR
\$POST	HHNR	HHNRAKT*	PERSNR
\$KIND	HHNR	HHNRAKT*	PERSNR
\$PEQUIV	HHNR	HHNRAKT*	PERSNR
\$HBRUTTO	HHNR	HHNRAKT*	-
\$H	HHNR	HHNRAKT*	-
\$HGEN	HHNR	HHNRAKT*	-
GHOST	HHNR	HHNRAKT*	-
YPBRUTTO	HHNR	PERSNR	ERHEBJ
PFLEGE	HHNR	PERSNR	ERHEBJ
BIOIMMIG	HHNR	PERSNR	ERHEBJ
ARTKALEN	HHNR	PERSNR	SPELLNR
EINKALEN	HHNR	PERSNR	SPELLNR
PBIOSPE	HHNR	PERSNR	SPELLNR
BIOMARSM	HHNR	PERSNR	SPELLNR
BIOMARSY	HHNR	PERSNR	SPELLNR
SOZKALEN	HHNR	HHNRAKT	SPELLNR
BIOBIRTH	HHNR	PERSNR	
BIOBRTHM	HHNR	PERSNR	
BIOAGE01	HHNR	PERSNR	
BIOAGE03	HHNR	PERSNR	
BIOAGE17	HHNR	PERSNR	
BIOTWIN	HHNR	PERSNR	
BIOPAREN	HHNR	PERSNR	
BIOSOC	HHNR	PERSNR	
BIORESID	HHNR	PERSNR	
BIOJOB	HHNR	PERSNR	

* also available as \$HHNR (\$=A, B, C, ..., wave)

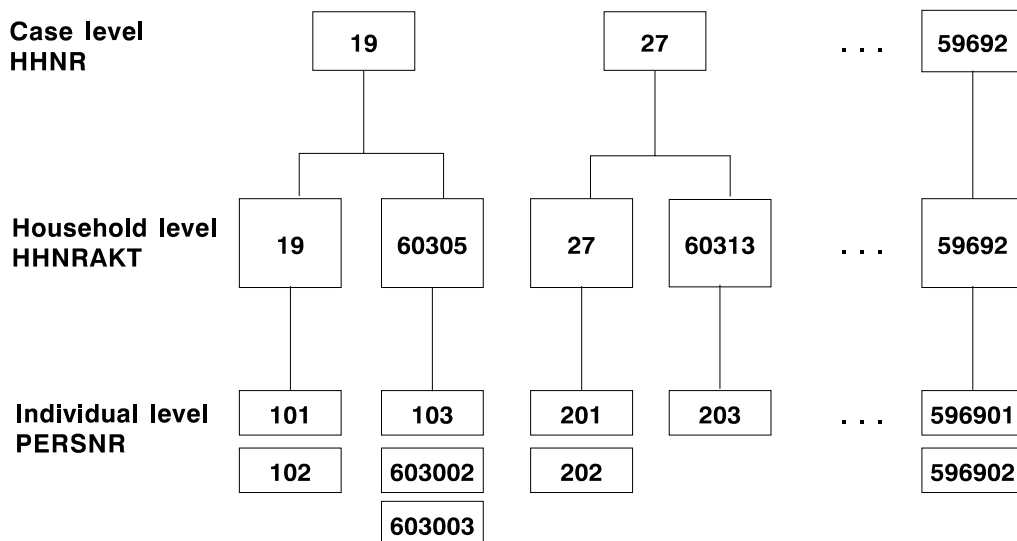
Linking data across time (2)

- Identifiers in the course of time

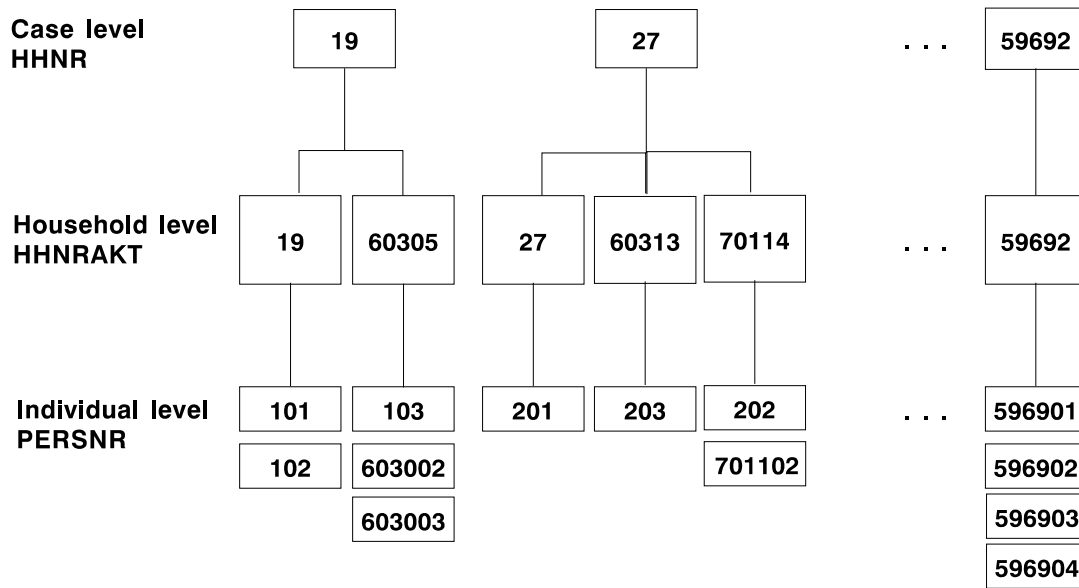
Wave 1 (1984)



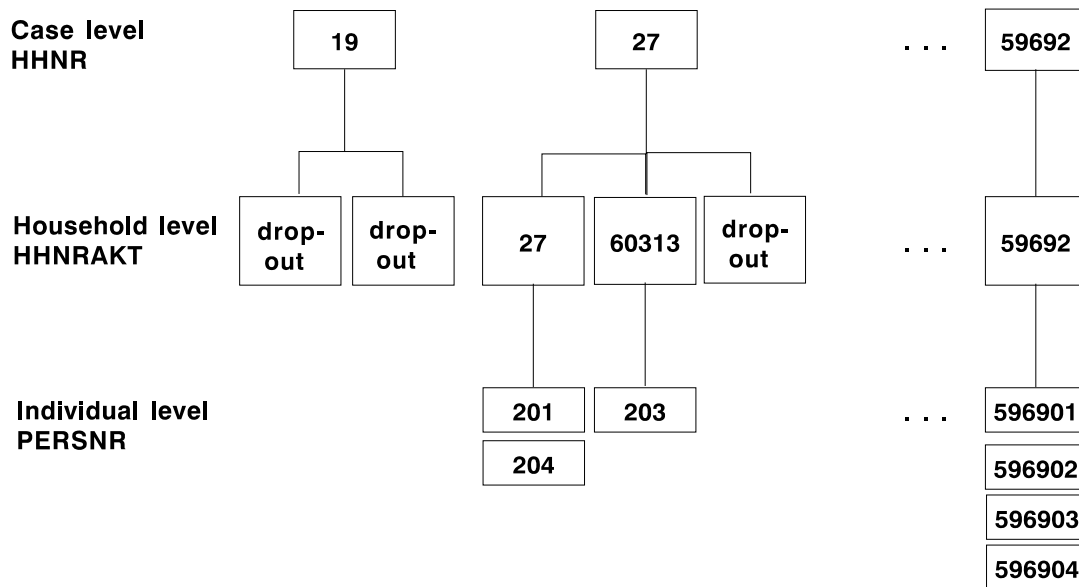
Wave 2 (1985)



Wave 5 (1988)



Wave 11 (1994)



Dealing with variables (1)

Principles for naming Survey Variables (up to 8 digits):

<u>digit</u>	<u>Meaning</u>
1	<u>Wave</u> (A=wave 1, B=wave 2,... according to West-Samples)
2	Differentiation according to unit of analysis: <u>P</u> =individual, <u>H</u> =household
3-4	<u>Number of question</u> in survey instrument (questionnaire)
5-6	<u>Number of item</u> in survey instrument (questionnaire)
5 or 7	Differentiation according to sample: <u>A</u> =Foreigners, <u>O</u> =East Germans
5	Differentiation according to green (<u>G</u>) and blue (<u>B</u>) questionnaire version for old and new households and persons, respectively
2 thru 8	<u>Text</u> for Variables in \$BRUTTO files and some occupation-related variables in \$P files

Examples:

AP04	Wave 1; Individual; Question 4
BH0502	Wave 2; Household; Question 5; Item 2
DP24G09	Wave 4; Individual; Question 24; Green version; Item 9
BIS88	Wave 2; 4 Digit Intl. Standard Classification of Occup. (ISCO88)
AP64A	Wave 1; Individual; Question 64; Sample B, Foreigners only

Exceptions:

- identifiers HHNR, PERSNR, HHNRAKT, SPELLNR, ERHEBJ
- generated and status variables in \$PGEN, \$HGEN and \$PEQUIV-files:

Examples:

- NATION95 Nationality 1995 (individual level, file LPGEN)
- PARTNR88 PERSNR of partner 1988 (individual level, file LPGEN)
- KTYPHH2 2-digit Household-Typology 1994 (household level, file KHGEN)
- FBAUJ Year of construction 1989 (household level, file FHGEN)
- I1110204 Household Post-Gov't Income - previous year, 2004 (individual level, file UPEQUIV)

Dealing with variables (2)

Declaring missing values

While identifiers like HHNR, PERSNR, HHNRAKT, SPELLNR, ERHEBJ are not allowed to be missing at all, survey variables might be missing for different purposes. A person can refuse to answer to a question (very often income-related questions) or just might not know an answer. Otherwise a question might not apply to a person or a household; e.g. the rent to be paid when the household is an owner-occupier.

The SOEP data differentiates three kinds of missing values:

<i>Code</i>	<i>Meaning</i>
-1	no answer / do not know
-2	does not apply
-3	after checking for plausibility a given value was found to be implausible and was finally deleted (thus, this code is to be interpreted like -1)

Important note on missing values:

The SOEP scientific use version (95% sample) might use a different coding of the missing values:

<i>Code 100% sample</i>	<i>Code 95% sample (old SAS-Version)</i>
-1	.A
-2	.B
-3	.C

Preparing data for analysis (1)

- *Cross-section data*

a) Single cross-section data (i=individuals 1, ... , n ; v=variables 1, ... , m)

	V ₁	.	.	.			V _m
i ₁							
.							
.							
.							
i _n							

b) Comparison of two cross-sectional datasets

	V ₁₁	.	.	.			V _{m1}
i ₁₁							
.							
.							
.							
i _{n1}							

	V ₁₂	.	.	.			V _{m2}
i ₁₂							
.							
.							
.							
i _{n2}							

c) Pooling of two cross-sectional datasets

	V _{1t}	.	.	.			V _{mt}	t
i ₁₁								1
.								.
.								.
.								.
i _{n1}								
i ₁₂								2
.								.
.								.
.								.
i _{n2}								

Preparing data for analysis (2)

- *Longitudinal data*

a) Complete case analysis with a balanced panel design

t_0	t_1	t_2	
			drop-outs
			successfully interviewed in all waves
			new respondents

not yet in the sample or not yet interviewed

b) Downstream model

t_0	t_1	t_2	
			drop-outs
			Successfully interviewed in all waves
			new respondents

not yet in the sample or not yet interviewed

c) Complete information analysis with an unbalanced panel design

t_0	t_1	t_2	
			drop-outs
			Successfully interviewed in all waves
			new respondents

not yet in the sample or not yet interviewed

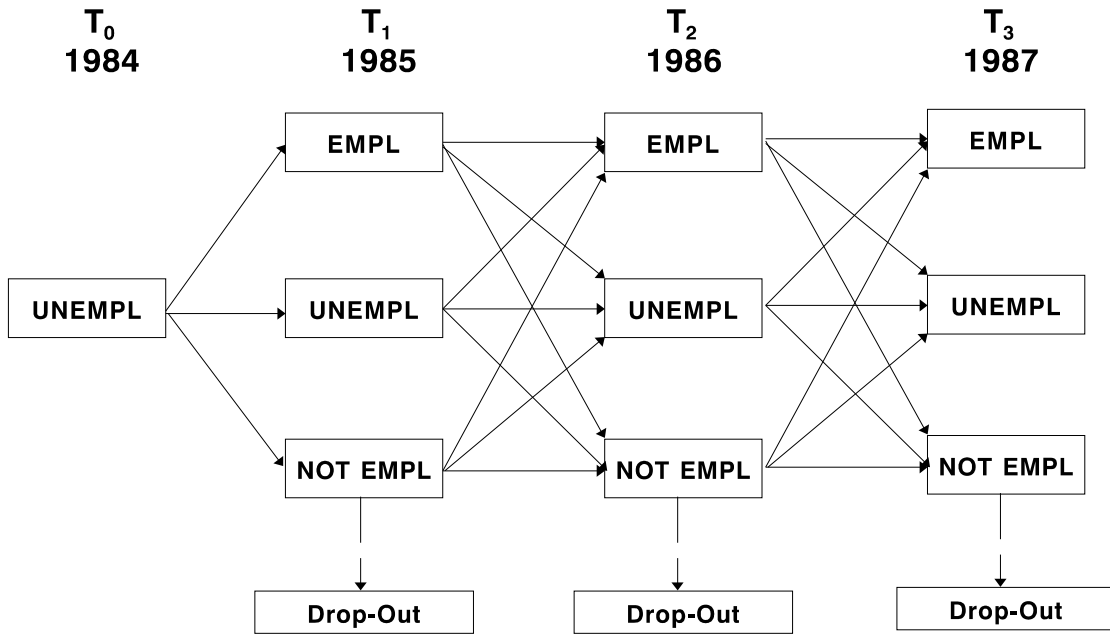
d) Pooling longitudinal data (two longitudinal datasets of two waves each)

t_0	t_1	t_2	
			drop-outs
			Successfully interviewed in all waves
			waves
			new respondents

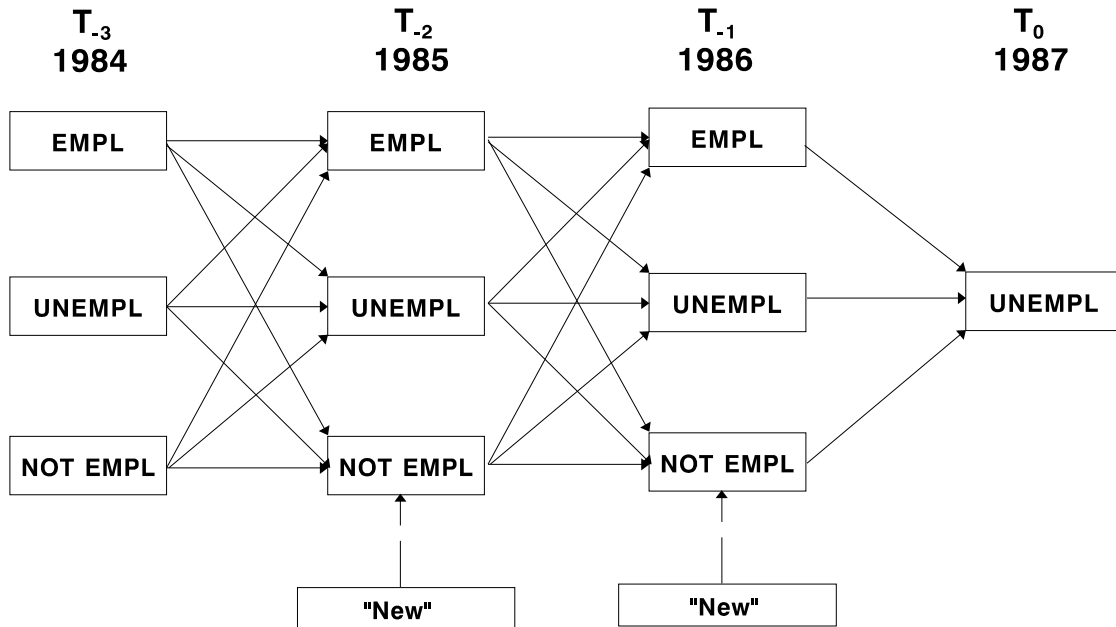
not yet in the sample or not yet interviewed

t_0	t_1
t_1	t_2

Downstream Model 1984 to three years later



Upstream Model 1987 to three years earlier



Preparing data for analysis (3)

- *Event and Spell data*

Although in the course of time the absolute number of observations (households and individuals) is almost steadily decreasing from a cross-sectional perspective, the cumulative number of events and/or spells covered with the entire data is increasing wave by wave.

Different SOEP files support the analysis of events ...

- Re-migration: More than 1,400 moves abroad can be identified up to 2003, with most of these events resulting from foreigners (in Sample B) returning to their home countries. These events are documented in the cumulative drop-out file YPBRUTTO.
- Deaths (mortality): About 2,100 observations are documented in the drop-out file YPBRUTTO up to 2003.
- Births (fertility): Using data from the biography questionnaire and the yearly information from the \$PBRUTTO files (which also covers changes in household composition due to demographic reasons) up to 2003, there are almost 30,000 births documented in the biography file BIOBIRTH, with about 18,500 of these persons being identified within the SOEP population.

... and spells

- The file PBIOSPE contains spell data on labor market status on a yearly basis, starting with the individual respondents age of 16.
- Examples (as of 2003)
 - about 16,000 spells of unemployment,
 - about 51,400 spells of full-time employment.
- The file ARTKALEN provides monthly spell data on 14 different kinds of labor-market involvement, defining begin, end, and censoring status of each spell of e.g. full-time work, part-time work, or unemployment.
- The file EINKALEN contains monthly spell data on periods during which individuals received different types of income (such as income from employment, pensions, unemployment benefits).
- The files BIOMARSM and BIOMARSY contain spell information on marital status of each respondent.
- BIOMARSM is on a monthly basis covering the period of time the respondent has been observed in the SOEP.
- BIOMARSY is on a yearly basis for the respondent's entire lifespan.
- Additional information on spells:
 - For each spell in the files PBIOSPE, ARTKALEN, EINKALEN there is a variable CENSOR, indicating whether the spell is
 - uncensored (begin and end are known from the SOEP data) or
 - censored:
 - left-censored: the begin of a given spell is unknown,
 - right-censored: the end of a given spell is unknown,
 - left and right-censored: begin and end of a given spell are unknown.
 - For each spell in the files BIOMARSM and BIOMARSY there is a variable REMARK giving information on the logical consistency of a spellsystem for a given person, i.e. indicating whether there are any implausible changes in marital status such as being widowed without ever having been married.

▪ *Principal structure of a spell-system in the SOEP*

with h=households 1, ... m
 i=individuals 1, ... , n
 s=spell 1, ... , o

HHNR	PERSNR	SPELLNR	SPELLTYP	BEGIN	END	CENSOR / REMARK
h ₁	i ₁₁	S ₁₁₁				
h ₁	i ₁₁	...				
h ₁	i ₁₁	S _{11o}				
h ₁				
h ₁	i _{1n}	S _{1n1}				
h ₁	i _{1n}	...				
h ₁	i _{1n}	S _{1no}				
...				
...				
...				
h _m	i _{m1}	S _{m11}				
h _m	i _{m1}	...				
h _m	i _{m1}	S _{m1o}				
h _m				
h _m	i _{mn}	S _{mn1}				
h _m	i _{mn}	...				
h _m	i _{mn}	S _{mno}				

Introduction to the German *Socio-Economic Panel* Study (SOEP)

- Part III: Extending the SOEP survey data -

(A) Cross-sectional files

- \$PGEN (Status variables and generated variables: individual level)
- \$HGEN (Status variables and generated variables: household level)
- \$PEQUIV (SOEP part of CNEF equivalent file: individual level w/ HH-info)
- \$PLUECKE (temporary drop-outs)

(B) The longitudinal files

- The meta-file PPFAD (individual level)
- The meta-file HPFAD (household level)
- Variables P/HSAMPLE and \$SAMPREG: How to correctly define cross-sectional populations for West and East Germany!
- PFLEGE (Persons needing care)

(C) Biographical and spell information in the SOEP

- General introduction to biographical data
- Files with biographical information on individual level:
 - \$PGEN and PPFAD
 - YPBRUTTO (drop-outs only, panel structure)
 - BIOBIRTH (women only)
 - BIOBRTHM (men only, since 2001)
 - BIOPAREN (all respondents)
 - BIOJOB (all respondents)
 - BIOAGE01 (mothers with newborn children, since 2003)
 - BIOAGE03 (mothers with 2-3 year old children, since 2005)
 - BIOAGE17 (16-17 year old first time respondents, since 2001)
 - BIOSOC (all respondents)
 - BIORESID (all respondents, since 1994)
 - BIOTWIN (multiples)
 - BIOIMMIG (immigrants only; panel structure)
- Files with spell data
 - BIOMARSY (individuals, marital history, yearly)
 - BIOMARSM (individuals, marital history, monthly)
 - PBIOSPE (individuals, occupation history, yearly)
 - ARTKALEN (individuals, occupation history, monthly)
 - EINKALEN (individuals, income receipt, monthly)
 - SOZKALEN (household, social assistance reciprocity)
- NEWSPELL: software to rearrange spell information

(D) Integration of data on West and East Germany in subsamples A, B and C

(A) Cross-sectional files

Status variables and generated variables in the SOEP

Status variables:

Problem:

- Some of the SOEP information is gathered in the course of the first interview, only.
- Some information is asked for in separate questions for “old” and “new” households and individuals, respectively.
- In some cases old respondents are asked for changes since last year’s interview only, while new respondents have to fill in the current status.
- Respondents in different subsamples might be asked for the same information in different questions.

Solution:

- In all these cases, the collected information is stored in different variables. In order to minimize computing efforts for the user, the SOEP provides yearly **status variables** on individual and household level, which integrate all of these information in a common variable showing the current status for all respondents. Thus, there is nothing else but a re-organisation of already existing data and there is almost no assumption or normative setting involved in the generating process.

Example: Status variable \$WOHNFL in household file \$HGEN

- The size of the house/apartment (in square meters) in which a household lives is asked in the first interview as well as after each residential move. An old household remaining at the old address is asked only if there were changes in the house/apartment due to reconstruction, etc. Thus, the 1994 information for a given household might have been gathered in 1984 or in any year since then. The generated variable \$WOHNFL carries on this old, but still valid, information until it has to be actualized due to a change. If there was a change, but the new information is missing, \$WOHNFL is defined as missing (Code A or code -1 depending on 95% or 100% version of the data).

Generated variables:

In addition to the above mentioned status variables the SOEP provides some generated variables for households and individuals, with the latter being often based on assumptions (see also the corresponding documentation on the CD-ROM).

Example: Generated variables PARTZ\$\$ and PARTNR\$\$ in individual file \$PGEN

- Each individual respondent is asked whether or not he/she lives with a partner (in a common household). Variables used to describe such a relationship in the SOEP-data are marital status, changes in marital status since last year, the relationship to the head of the household, and information taken from the family and marriage biography.
While the variable PARTZ\$\$ defines the “quality” of the partnership (legally married or cohabiting), the variable PARTNR\$\$ contains the partner’s unique individual identifier (variable PERSNR). This variable allows researchers to merge individual characteristics of both persons in a partnership.

▪ **Cross-Sectional Files (1)**

**List of variables in the cross-sectional files \$PGEN
(Example: Wave Q 2000, individual level, status and generated variables)**

<i>Variable Name</i>	<i>Meaning</i>	<i>ID /Status /Generated</i>
HHNR	original household identifier (case) from wave 1	ID
HHNRAKT	current household identifier	ID
PERSNR	unique individual identifier	ID
ERWTYP00	employment status 2000	G
ERLJOB00	working in the original job?	S
BETR00	firm size	S
OEFFD00	public sector	S
AUSB00	educational requirements of job	S
PARTZ00	kind of relationship to partner	G
PARTNR00	unique individual identifier of partner	G
NATION00	nationality	S
QPSBIL	highest school degree received	S
QPBBIL01	highest occupational degree received	S
QPBBIL02	university degree	S
QPBBIL03	no occupational degree	S
QPSBILA	highest school degree received abroad (Sample B)	S
QPBBILA	highest occ. degree received abroad (Sample B)	S
QPSBILO	highest school degree (Sample C)	S
QPBBILO	highest occ. degree received (Sample C)	S
QFAMSTD	marital status	S
QBILZEIT	institutional years necessary to receive current degree of education	G
QERWZEIT	years with current employer	G
QTATZEIT	effective working time in hours per week	S
QVEBZEIT	contracted working time in hours per week	S
QUEBSTD	overtime last week	S/G
LFS00	labor force status	S/G
IS8800	ISCO88-4-digit	S
ISEI00	ISEI-Status88 according to GANZEBOOM	S
MPS00	Magnitude Prestige Scale (based on KLAS94)	S
NACE00	NACE industry codes	S
SIOPS00	TREIMAN Standard Int. Occ. Prestige Score	S
EGP00	ERIKSON and GORLDTORPE Class Categories	S
KLAS00	Classification of occupation (Statistical Office)	S
AUTONO00	Autonomy of job	G
ISCED00	International Classification of Education	G
CASMIN00	CASMIN Classification	G
STIB00	Occupational position	G
MONTH00	month of interview	G
MODE00	mode of interview	G

... contd.

<i>Variable Name</i>	<i>Meaning</i>	<i>ID /Status /Generated</i>
LABGRO00	monthly gross labor income in EURO	G
LABNET00	monthly net labor income in EURO	G
IMPGRO00	monthly gross labor income (imputation flag)	G
IMPNET00	monthly gross labor income (imputation flag)	G
ALLBET00	Firm size (longitudinally harmonized) '	G
EMPLST00	Employment Status	G
EXPFT00	Labor Market Experience, years in full-time	G
EXPPT00	Labor Market Experience, years in part-time	G
EXPUE00	Labor Market Experience, years in unemployment	G

▪ **Cross-Sectional Files (2)**

**List of variables in the cross-sectional file \$HGEN
(Example: Wave Q 2000, household level, status- and generated variables)**

<i>Variable Name</i>	<i>Meaning</i>	<i>ID /Status /Generated</i>
HHNR	original household identifier (case) from wave 1	ID
HHNRAKT	current household identifier	ID
QEINZUG	year moved into house or apartment	S
QBAUJ	year of construction (categorized)	S
QRENOV	degree to which dwelling needs repair	S
QWOHNFL	size of housing unit in square-meter	S
QWOHNR	no. of rooms in dwelling	S
QWGURT	evaluation of apartment-size	S
QAUS1	kitchen	S
QAUS2	bath, shower	S
QAUS3	toilet in dwelling	S
QAUS4	central heating	S
QAUS5	balcony, porch	S
QAUS6	cellar	S
QAUS7	garden	S
QAUS8	hot water, boiler	S
QAUS9	telephone	S
QEIGEN	owner-occupier or tenant	S
QERWERB	type of acquisition	S
QFOERD	support by public loans	S
QMIETE	monthly rent in DM	S
QNOMIET	don't have to pay rent	S
QMURT	evaluation of rent to be paid	S
QSOZIAL	social housing	S
QBILLIG	rent reduced by owner ?	S
QKOSTEN	monthly cost for hot water, heating	S
QTYPHH1	household typology (1-digit)	G
QTYPHH2	household typology (2-digit)	G
QMIETEG	monthly rent in DM	G
QHEIZG	monthly costs for heating+hot water	G
HMONTH00	month of household interview	G
HMODE00	mode of household interview	G
HINC00	monthly net household income (EURO)	S
AHINC00	Adjusted monthly net household income (EURO)	G

- **Cross-Sectional Files (3)**

List of variables in the cross-sectional file \$PEQUIV

Most of this data is taken from the Cross-National-Equivalent data file (CNEF), produced by Cornell University in cooperation with the SOEP group. In contrast to the original CNEF-data which is based on the 95% scientific use file of SOEP, the \$PEQUIV-files include the full 100%-sample. For extensive documentation of the CNEFcf. <http://www.human.cornell.edu/pam/gsoep/equivfil.cfm>

Burkhauser, Richard V, Barbara A. Butrica, Mary C. Daly and Dean R. Lillard 2001: The Cross-National Equivalent File: A product of cross-national research. In: Becker, Irene, Ott, Notburga and Rolf, Gabriele (Eds.) Soziale Sicherung in einer dynamischen Gesellschaft (Social Insurance in a Dynamic Society). Festschrift für Richard Hauser zum 65. Geburtstag (Papers in Honor of the 65th Birthday of Richard Hauser), Campus, Frankfurt/New York

- Waves A (1984) - V (2005)
- Single income components and household aggregated annual income measures (as of previous year) derived from household questionnaire as well as from the individual questionnaire (incl. the income calendar running from January through December as of previous year)
- Population for \$PEQUIV is made up by all members of households with a successful interview (i.e., persons with \$NETTO-codes 1 to 5 in the file PPFAD and \$HNETTO-code 1 in the file HPFAD).
- Income data is missing for Sample C in 1990 and 1991 (first 2 waves of East German sample)
- All income measures are given in Euro and imputed in case of item-non-response (for technical details see documentation Grabka & Frick (2003): http://www.diw.de/deutsch/produkte/publikationen/materialien/docs/papers/diw_rn03-10-29.pdf)

List of variables in the cross-sectional file \$PEQUIV (Example: Wave T 2003, individual level)

Identifiers and Demographics

HHNR	Original Household Number
HHNRAKT	Current Wave HH Number (=THHNR)
THHNR	Current Wave HH Number (=THHNR)
PERSNR	Never Changing Person ID
X11101LL	Person Identification Number
D11102LL	Gender of Individual
X11104LL	Subsample Identifier
X1110203	HH Identification Number
X1110303	Individual in HH at Survey
X1110503	Individual responded to Survey
D1110103	Age of Individual
D1110303	Race of HH Head
D1110403	Marital Status of Individual
D1110503	Relationship to HH Head
D1110603	Number of Persons in HH
D1110703	Number of Children in HH
D1110803	Education With Respect to High School
D1110903	Number of Years of Education
D1111003	Disability Status of Individual
D1111103	Satisfaction With Health

Employment variables

E1110103	Annual Work Hours of Individual
E1110203	Employment Status of Individual
E1110303	Employment Level of Individual
E1110403	Primary Activity of Individual
E1110503	Occupation of Individual
E1110603	1 Digit Industry Code of Individual
E1110703	2 Digit Industry Code of Individual
E1120103	Impute Annual Work Hours of Individual

Household composition

H1110103	Number of hh members age 0-14
H1110203	Number of hh members age 15-18
H1110303	Number of hh members age 0-1
H1110403	Number of hh members age 2-4
H1110503	Number of hh members age 5-7
H1110603	Number of hh members age 8-10
H1110703	Number of hh members age 11-12
H1110803	Number of hh members age 13-15
H1110903	Number of hh members age 16-18
H1111003	No. hh members 19 and above or 16-18,ind
H1111103	Indicator - Head in HH
H1111203	Indicator-wife in HH

CNEF- Annual Income Variables

I1110103	HH Pre-Government Income
I1110203	HH Post-Government Income
I1110303	HH Labor Income
I1110403	HH Income From Asset Flows
I1110503	HH Imputed Rent
I1110603	HH Private Transfers
I1110703	HH Public Transfers

I1110803	HH Social Security Pensions
I1110903	Total HH Taxes
I1111003	Individual Labor Earnings
I1111103	HH Federal Taxes
I1111203	HH Social Security Taxes
I1111303	HH Post-Government Income (TAXSIM)
I1111403	Total HH Taxes (TAXSIM)
I1111503	HH State Taxes (TAXSIM)
I1111603	HH Federal Taxes (TAXSIM)
I1111703	HH Private Retirement Income
I1111803	Household Windfall Income
I1120103	Impute HH Pre-Government Income
I1120203	Impute HH Post-Government Income
I1120303	Impute HH Labour Income
I1120403	Impute HH Income From Asset Flows
I1120603	Impute HH Private Transfers
I1120703	Impute HH Public Transfers
I1120803	Impute HH Social Security Pensions
I1121003	Impute Individual Labor Earnings
I1121703	Impute HH Private Retirement Income
I1121803	Impute Household Windfall Income
I1120503	Impute HH Imputed Rental Value
I1120903	Impute Total HH Taxes

Weights

W1110103	X-Sectional Weight - Respondent Individual
W1110203	HH Weight
W1110303	Longitudinal Weight - Respondent Individual
W1110503	Individual Weight - Immigrant Sample
W1110603	HH Weight - Immigrant Sample
W1110703	X-Sectional Weight - Enumerated Individual
W1110803	Longitudinal Weight - Enumerated Individual
W1110903	Population Factor for w11103\$\$
W1111003	Population Factor for w11107\$\$
W1111103	Population Factor for w11108\$\$
W1110403	Population Factor for W11101\$\$

Macro Indicators

Y1110103	Consumer Price Index
L1110103	State of Residence
L1110203	Region

Additional individual income variables

IJOB103	Wages,Salary from main job
IJOB203	Income from secondary employment
ISELF03	Income from self-employment
IOLDY03	old-age,disability and civil serv. pensions
IWIDY03	widows andor orphans pension
IUNBY03	Unemployment benefit
IUNAY03	Unemployment assistance
ISUBY03	Subsistence allowance

IERET03	Old-age transition benefit
IMATY03	Maternity benefit
ISTUY03	Student grants
IMILT03	Militarycommunity service pay
IALIM03	Alimony
IELSE03	Private Transfers received
ICOMP03	Company pension (surviving dependants c.p.)
IPRVP03	Private pension (old-age,accid.,disability)
I13LY03	13th monthly salary
I14LY03	14th monthly salary
IXMAS03	Christmas bonus
IHOLY03	Vacation bonus
IGRAY03	Profit-sharing
IOTHY03	Other bonuses
IGRV103	Retirement pay: stat. pension insurance
IGRV203	Widows pension: stat pension insurance
RENTY03	Income from rental and leasing
OPERY03	Operation, maintenance costs
DIVDY03	Interest, dividend income
CHSPT03	Child allowance
HOUSE03	Housing benefit
NURSH03	Compulsory long term care insurance
SUBST03	Social assistance(living expenses etc)
SPHLP03	Social assistance f. spec. circumstances
HSUP03	Housing support for owner-occupiers
FJOB103	Imp.flag:Wages,Salary from main job
FJOB203	Imp.flag:Income from secondary job
FSELF03	Imp.flag:Income from self-employment
FUNBY03	Imp.flag:Unemployment benefit
FOLDY03	Imp.flag:old-age,civil serv. pensions
FWIDY03	Imp.flag:widowsorphans pension
FUNAY03	Imp.flag:Unemployment assistance
FSUBY03	Imp.flag:Subsistence allowance
FERET03	Imp.flag:Old-age transition benefit
FMATY03	Imp.flag:Maternity benefit
FSTUY03	Imp.flag:Student grants
FMILT03	Imp.flag:Militarycommunity service pay
FALIM03	Imp.flag:Alimony
FELSE03	Imp.flag:Private Transfers received
FCOMP03	Imp.flag:Company pension
FPRVP03	Imp.flag:Private pension(old-age,accid.)
F13LY03	Imp.flag:13th monthly salary
F14LY03	Imp.flag:14th monthly salary
FXMAS03	Imp.flag:Christmas bonus
FHOLY03	Imp.flag:Vacation bonus
FGRAY03	Imp.flag:Profit-sharing
FOTHY03	Imp.flag:Other bonuses
FGRV103	Imp.flag:retirement pay from stat.insurance
FGRV203	Imp.flag:widows pension from stat.insurance
FRENTY03	Imp.flag:Income from rental and leasing
FOPERY03	Imp.flag:Operation, maintenance costs
FDIVDY03	Imp.flag:Interest, dividend income
FCHSPT03	Imp.flag:Child allowance
FHOUSE03	Imp.flag:Housing benefit
FNURSH03	Imp.flag:Compuls. long term care

	insurance
FSUBST03	Imp.flag:Social assist.(living expenses ..)
FSPHLP03	Imp.flag:Soc. assist. for spec. circumstan.
FHSUP03	Imp.flag:Housing support f. owner-occupiers
ISMP103	Social miners insurance pension
ICIV103	Civil servant pension
IWAR103	War victim pension
IAGR103	Farmer Pension
IGUV103	Statutory accident insurance
IVBL103	Supplementary benefits for civil servants
ICOM103	Company pension
IPRV103	Private pension
ISON103	Other pension
ISMP203	Widows social miners insurance pension
ICIV203	Widows civil servant pension
IWAR203	Widows war victim pension
IAGR203	Widows farmer Pension
IGUV203	Widows statutory accident insurance
IVBL203	Widows supplement. benefits(civil servants)
ICOM203	Widows company pension
ISON203	Other widows pension
IPRV203	Widows private pension

Health related variables

M1110103	Overnight hosp stay
M1110203	Inpatient nights in hosp
M1110303	Work accident required treatment
M1110403	Frequency of sport or exercise
M1110503	Have had stroke
M1110603	High blood pressurecirculation problems
M1110703	Have or had diabetes
M1110803	Have or had cancer
M1110903	Psychiatric problems
M1111003	Arthritis
M1111103	Angina or heart condition
M1111203	Difficulties breathing
M1111303	Have trouble climbing stairs
M1111403	Need help or have difficulty bathing alone
M1111503	Dressing difficult alone
M1111603	Difficultyneed help getting in/out bed
M1111703	Need help with shopping
M1111803	Walk 10+ min alone difficult
M1111903	Housework difficult alone
M1112003	Health limits kneeling
M1112103	Health limits vigorous activities
M1112203	Body height
M1112303	Body weight
M1112403	Disability Status of Individual
M1112503	Satisfaction with Health
M1112603	Current Self-Rated Health Status
M1112703	Number of annual doctor visits

- **Cross-Sectional Files (4)**

The temporary drop-out files \$PLUECKE

Temporary drop-outs (“gaps”) can cause problems for longitudinal analyses. This is especially true for the employment and income data stored in the spell-files ARTKALEN and EINKALEN. That is why the SOEP tries to fill in at least some of the central missing information. Persons who take part in the survey after such a temporary unit non-response are asked to complete:

- the individual questionnaire for the current wave, plus
- a small questionnaire covering information of the year in which the drop-out occurred. This includes questions on:
 - job-related changes,
 - calendar of occupation and income (is included in the spell-data),
 - education and qualification.

Note:

- This additional data is stored in the cross-sectional files \$PLUECKE.
- The variable names correspond to those in the same wave’s \$P file.
- Persons with a completed “gap”-questionnaire are marked in the corresponding \$NETTO variable in PPFAD with the code 4.

Exception: Persons who did not live in Germany at the time of last year’s interview are coded with \$NETTO = -3.

(B) The longitudinal meta-files

Longitudinal Meta Data (1)

Individuals in PPFAD

This file includes all members of all households ever contacted in the SOEP including respondents, children, and even those who never gave an interview. For each person it includes the household identifiers for each wave (\$HHNR) as well as wave-specific variables concerning the survey status (\$NETTO). Thus, PPFAD's central function is to facilitate the definition of longitudinal populations.

Example for \$NETTO variables: **QNETTO** Survey status 2000

<i>Label</i>	<i>Value</i>	<i>Frequency</i>	
		100%	95%
in YPBRUTTO	0	217	206
in QP	1	24,586	23,341
in QKIND	2	6,659	6,295
in QPBRUTTO	3	2,640	2,475
in QPLUECKE	4	190	179
Total		34,292	32,496
Missing*	-2 / .B	16,147	15,347
Total in PPFAD (as of wave 2001)		50,439	47,843

* Individuals who did not yet enter the survey or who already left.

Additionally, PPFAD contains the longitudinally checked information on gender and year of birth for each individual. For the sake of consistency this information should be used in longitudinal analyses, since it can not be taken for granted, that the corresponding yearly or cross-sectional information is perfectly stable.

SEX gender (longitudinally verified)

<i>Label</i>	<i>Value</i>	<i>Frequency</i>	
		100%	95%
male	1	24,987	23,715
female	2	25,441	24,117
Total		50,428	47,832
Missing	-1 / .A	11	11
Total in PPFAD (as of wave 2001)		50,439	47,843

List of variables in PPFAD (as of 2005)

<i>Variable Name</i>	<i>Meaning</i>
HHNR	original household identifier (case) from wave 1
PERSNR	unique individual identifier
PSAMPLE	sample identifier
SEX	gender (longitudinally verified)
GEBJAHR	year of birth (4 digit) longitudinally verified
GEBMONAT	month of birth
AHHNR	household identifier 1984
BHHNR	household identifier 1985
CHHNR	household identifier 1986
...	...
VHHNR	household identifier 2005
ANETTO	survey status 1984
BNETTO	survey status 1985
CNETTO	survey status 1986
...	...
VNETTO	survey status 2005
EINTRITT	year in which individual entered the survey (4 digit)
ERSTBEFR	year in which first individual interview was conducted (4 digit)
AUSTRITT	year in which individual left the survey (4 digit)
LETZTBEF	year in which last individual interview was conducted (4 digit)
TODJAHR	year of death (4-digit)
TODINFO	source of information to compute year of death
IMMIYEAR	year of first immigration to Germany after 1948 (4 digit)
GERMBORN	born in Germany or immigration prior to 1949
CORIGIN	country of origin
LOC1989	Where did you live in 1989 ?
LCASEMAT	HHNR of case-match 1995
MCASEMAT	HHNR of case-match 1996
...	...
VCASEMAT	HHNR of case-match 2005
GSAMPREG	Region in which household lives (West or East Germany) 1990
HSAMPREG	Region in which household lives (West or East Germany) 1991
...	...
VSAMPREG	Region in which household lives (West or East Germany) 2005
APOP	Population 1984
BPOP	Population 1985
...	...
VPOP	Population 2005

Records and Variables Contained in PPFAD (1) Example based on data as of 2001

HHNR	PERSNR	PSAMPLE	SEX	GEBJAHR	TODJAHR	IMMIYEAR	GERMERN	EINTRITT	ERSTBEFR	AUSTRITT	LETZTBEF
27	201	1	2	1926	-2	-2	1	1984	1984	2001	2001
27	202	1	2	1956	-1	-2	1	1984	1985	1988	1987
27	203	1	1	1960	-2	-2	1	1984	1985	2001	2001
27	204	1	1	-1	1995	-2	1	1990	-2	1995	-2
27	701102	1	1	1956	-1	-2	1	1987	-2	1988	-2
45799	457901	2	1	1932	-1	1969	0	1984	1984	1986	1985
45799	457902	2	2	1936	-1	1970	0	1984	1984	1986	1985
45799	457903	2	1	1966	-1	1970	0	1984	1984	1986	1985
45799	457904	2	1	1974	-1	-2	1	1984	-2	1986	-2
500046	5000401	3	1	1967	-2	-2	1	1990	1990	1997	1996
500046	5000402	3	1	1936	-2	-2	1	1990	1990	2001	2001
500046	5000403	3	2	1939	-2	-2	1	1990	1991	2001	2001
500046	5503902	3	2	1967	-2	-2	1	1993	1993	1997	1996
500046	5503903	3	2	1992	-2	-2	1	1993	-2	1997	-2
700010	7000101	4	1	1951	-2	-2	1	1994	1994	2001	2001
700010	7000102	4	2	1962	-2	-2	1	1994	1994	2001	2001
720054	7200501	4	1	1933	1996	1990	0	1995	1995	1996	1995
720054	7200502	4	2	1937	-2	1994	0	1995	1995	1999	1998
200034	2000301	5	2	1918	-2	-2	1	1998	1998	2001	2000
200034	2000302	5	1	1949	-2	-2	1	1998	1998	2001	2001
200034	2000303	5	1	1943	-2	-2	1	2001	2001	2001	2001
250015	2500101	6	1	1940	-2	-2	1	2000	2000	2001	2001
250015	2500102	6	1	1925	2001	-2	1	2000	2000	2000	2000

Records and Variables Contained in PPFAD (2)

HHNR	PERSNR	ANETTO	BNETTO	CNETTO	DNETTO	ENETTO	FNETTO	GNETTO	HNETTO	INETTO	JNETTO	KNETTO	..	PNETTO	QNETTO
27	201	1	1	1	1	1	1	1	1	1	1	1		1	1
27	202	3	1	1	1	3	-2	-2	-2	-2	-2	-2		-2	-2
27	203	3	1	1	1	1	1	1	1	1	1	1		1	1
27	204	-2	-2	-2	-2	-2	-2	3	3	3	3	3		-2	-2
27	701102	-2	-2	-2	3	3	-2	-2	-2	-2	-2	-2		-2	-2
45799	457901	1	1	3	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
45799	457902	1	1	3	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
45799	457903	1	1	3	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
45799	457904	2	2	3	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
500046	5000401	-2	-2	-2	-2	-2	-2	1	1	1	1	1		-2	-2
500046	5000402	-2	-2	-2	-2	-2	-2	1	1	1	1	1		1	1
500046	5000403	-2	-2	-2	-2	-2	-2	3	1	1	1	1		1	1
500046	5503902	-2	-2	-2	-2	-2	-2	-2	-2	-2	1	1		-2	-2
500046	5503903	-2	-2	-2	-2	-2	-2	-2	-2	-2	2	2		-2	-2
700010	7000101	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	1		1	1
700010	7000102	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	1		1	1
720054	7200501	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
720054	7200502	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		3	-2
200034	2000301	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		1	1
200034	2000302	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		1	1
200034	2000303	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		-2	-2
250015	2500101	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		-2	1
250015	2500102	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2		-2	1

Records and Variables Contained in PPFAD (3)

HHNR	PERSNR	AHHNR	BHHNR	CHHNR	DHHNR	EHHNR	FHHNR	GHHNR	HHHNR	IHHNR	JHHNR	KHHNR	PHHNR	QHHNR
27	201	27	27	27	27	27	27	27	27	27	27	27	27	27
27	202	27	27	27	70114	70114	-2	-2	-2	-2	-2	-2	-2	-2
27	203	27	60313	60313	60313	60313	60313	60313	60313	60313	60313	60313	60313	60313
27	204	-2	-2	-2	-2	-2	-2	27	27	27	27	27	-2	-2
27	701102	-2	-2	-2	70114	70114	-2	-2	-2	-2	-2	-2	-2	-2
45799	457901	45799	45799	45799	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
45799	457902	45799	45799	45799	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
45799	457903	45799	45799	45799	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
45799	457904	45799	45799	45799	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
500046	5000401	-2	-2	-2	-2	-2	-2	500046	500046	500046	550396	550396	-2	-2
500046	5000402	-2	-2	-2	-2	-2	-2	500046	500046	500046	500046	500046	500046	500046
500046	5000403	-2	-2	-2	-2	-2	-2	500046	500046	500046	500046	500046	500046	500046
500046	5503902	-2	-2	-2	-2	-2	-2	-2	-2	-2	550396	550396	-2	-2
500046	5503903	-2	-2	-2	-2	-2	-2	-2	-2	-2	550396	550396	-2	-2
700010	7000101	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	700010	700010	700010
700010	7000102	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	700010	700010	700010
720054	7200501	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
720054	7200502	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	720054	-2
200034	2000301	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	200034	200034
200034	2000302	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	200034	200034
200034	2000303	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
250015	2500101	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	250015
250015	2500102	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	250015

▪ **Longitudinal Meta Data (2)**

List of variables in household file HPFAD (as of 2004)

<i>Variable Name</i>	<i>Meaning</i>
HHNR	original household identifier (case) from wave 1
HHNRAKT	unique household identifier
HSAMPLE	sample identifier
AHHNR	household identifier 1984
BHHNR	household identifier 1985
CHHNR	household identifier 1986
...	
VHHNR	household identifier 2005
AHNETTO	survey status 1984
BHNETTO	survey status 1985
CHNETTO	survey status 1986
...	
VHNETTO	survey status 2005
GSAMPREG	Region in which household lives (West or East Germany) 1990
HSAMPREG	Region in which household lives (West or East Germany) 1991
...	
VSAMPREG	Region in which household lives (West or East Germany) 2005
...	
AHPOP	Population 1984
BHPOP	Population 1985
...	
VHPOP	Population 2005

▪ **Longitudinal Meta Data (3)**

The relationship of (P-/H-) SAMPLE and \$SAMPREG up to wave V, 2005

How to correctly define cross-sectional population (here: interviewed respondents plus children, i.e. \$NETTO=1 or 2) for West-and East Germany!

\$SAMPREG	Sample A	Sample B	Sample C	Sample D	Sample E	Sample F	Sample G	Total
1990 (Wave G)								
West G.	8717	3493	-	-	-			12210
East G.	-	-	6044	-	-			6044
1991 (Wave H)								
West G.	8667	3494	44	-	-			12205
East G.	0	0	5639	-	-			5639
1992 (Wave I)								
West G.	8526	3416	138	-	-			12080
East G.	2	0	5347	-	-			5349
1993 (Wave J)								
West G.	8450	3338	186	-	-			11974
East G.	8	0	5090	-	-			5098
1994 (Wave K)								
West G.	8336	3187	227	719	-			12469
East G.	11	0	4954	-	-			4962
1995 (Wave L)								
West G.	8254	2992	280	1592	-			13118
East G.	23	2	4781	3	-			4809
1996 (Wave M)								
West G.	8111	2896	294	1479	-			12780
East G.	27	2	4682	13	-			4724
1997 (Wave N)								
West G.	8009	2794	311	1407	-			12521
East G.	30	2	4550	23	-			4605
1998 (Wave O)								
West G.	7760	2618	294	1269	1959			13900
East G.	39	2	4373	19	417			4850
1999 (Wave P)								
West G.	7568	2519	326	1190	1663			13266
East G.	41	1	4267	23	372			4704
2000 (Wave Q)								
West G.	7311	2392	436	1149	1566	11275		24039
East G.	49	0	4167	27	355	2608		7206
2001 (Wave R)								
West G.	7165	2355	364	1077	1479	9301		21741
East G.	56	0	4002	27	333	2209		6627
2002 (Wave S)								
West G.	6946	2210	359	1032	1371	8502	2986	23406
East G.	62	0	3813	28	310	2036	378	6627
2003 (Wave T)								
West G.	6839	2116	394	1036	1325	7928	2247	21885
East G.	60	1	3740	26	290	1986	282	6385
2004 (Wave U)								
West G.	6655	2030	441	977	1292	7563	2151	22109
East G.	69	1	3649	26	291	1890	269	6195
2005 (Wave V)								
West G.	6393	1957	437	944	1230	7124	1978	20063
East G.	75	2	3494	26	282	1834	252	5965

Basis: successfully interviewed Households in 2002 (SHNETTO=1)

100% Sample

Sample	Region 2002 (SSAMPREG)		Total
	West-Germany	East Germany	
A	3,094	29	3,123
B	766	0	766
C	179	1,639	1,818
D	389	13	402
E	626	147	773
F	3,638	948	4,586
G	1,097	127	1,224
Total	9,789	2,903	12,692

95% Sample

Sample	Region 2002 (SSAMPREG)		Total
	West-Germany	East Germany	
A	2,929	26	2,955
B	731	0	731
C	172	1,557	1,729
D	372	13	385
E	596	139	735
F	3,461	896	4,357
G	1,043	120	1,163
Total	9,304	2,751	12,055

Basis: successfully interviewed Persons in 2002 (SNETTO=1)

100% Sample

Sample	Region 2002 (SSAMPREG)		Total
	West-Germany	East Germany	
A	5,532	45	5,577
B	1,598		1,598
C	289	3177	3,466
D	755	25	780
E	1,112	261	1,373
F	6,692	1735	8,427
G	2,375	296	2,671
Total	18,353	5539	23,892

95% Sample

Sample	Region 2002 (SSAMPREG)		Total
	West-Germany	East Germany	
A	5,245	42	5,287
B	1,527	0	1,527
C	273	3,021	3,294
D	724	25	749
E	1,059	246	1,305
F	6,348	1,624	7,972
G	2,256	280	2,536
Total	17,432	5,238	22,670

- **Longitudinal Data (4)**

PFLEGE: Persons needing care

- Since wave B (1985) the SOEP household questionnaire includes questions on household members in need of care. In order to support analyses on an individual level, this information has been restructured and stored in the cumulative file PFLEGE.
- For any person mentioned in the household questionnaire as needing care in a given year, now a single record in the file PFLEGE is provided. Since a person might show up in PFLEGE more than once, a second Sort-ID ERHEBJ indicates the year in which this person has been mentioned.
- Additional variables describe the intensity of care necessary (variables MAXGRAD and MULTGRAD) as well as which person provides the care (variable WERPFLGT, available for the years 1985 to 1990, only).

HHNR	Household identifier (case) of first wave
PERSNR	Unique individual identifier
ERHEBJ	Survey Year
MAXGRAD	Maximum of care intensity (1=lowest ... 5=highest intensity)
MULTGRAD	Multiple responses for different categories of care needs (1. digit = lowest ... 5. digit = highest intensity)
WERPLGT	Person(s) providing care (multiple responses possible: 1. digit = community nurse / social worker 2. digit = friends 3. digit = neighbors 4. digit = relatives in the household 5. digit = relatives outside the household 6. digit = can be anybody)

HHNR	PERSNR	ERHEBJ	MAXGRAD	MULTGRAD	WERPFLGT
43	401	1985	1	10000	100000
43	401	1986	3	100	10000
124	1201	1985	1	10000	10
213	2102	1993	1	10000	-2
213	2102	1994	1	10000	-2
213	2102	1995	5	10001	-2
213	2102	1996	1	10000	-2
779	7702	1991	-1	-1	-2
779	7702	1993	1	10000	-2
779	7702	1994	1	10000	-2
779	7702	1995	2	11000	-2
779	7702	1996	2	11000	-2
779	7702	1997	4	11010	-2
779	7702	1998	1	10000	-2

(C) Biographical information in the SOEP

see extended documentation:

Frick, Joachim R. and Thorsten Schneider (Eds.): Biography and Life History Data in the German Socio-Economic Panel (updated to Wave V, 2005), DIW Berlin.

<http://www.diw.de/deutsch/sop/service/doku/docs/bio.pdf>

Content

1. General Introduction
2. Biographical Information in the Meta File PPFAD (Month of Birth, Year of Death, Immigration Variables, Living in East or West Germany in 1989)
3. Activity Biography in the File PBIOSPE
4. BIOJOB: Detailed Information on First and Last Job
5. The Biography of Family Status and the Generated Current Family Status (BIOMARSY, BIOMARSM and \$FAMSTD)
6. BIOBIRTH – A Data Set on the Birth Biography of Female Respondents
7. BIOBRTHM – The Birth Biography of Male Respondents in the SOEP
8. BIOTWIN – Information on Twins in the SOEP
9. BIOIMMIG: Generated and Status Variables from SOEP for Foreigners and Migrants
10. BIOPAREN: Biography Information for the Parents of SOEP-Respondents
11. Retrospective Data on Youth and Socialisation Stored in BIOSOC
12. BIORESID: Variables On Occupancy and Second Residence
13. The Youth Questionnaire and the Corresponding Data Set BIOAGE17
14. BIOAGE01: Generated Variables from the “Mother and Child Questionnaire” (newborn children)
15. BIOAGE03: Generated Variables from the „Infant“ Questionnaire (2-3 year old children)

C.1 General introduction

1. In general, each SOEP respondent is asked to fill in additional questions covering biographical background information (Questionnaire “Life History”). Principally, this is done in the course of the first interview covering the following topics:

- occupational biography for ages 15 to 65
- family and marital biography
- social origin and first job
- migration biography

In 2000, a separate “Youth” questionnaire has been designed focusing on questions specifically targeted at youth/adolescence issues (to be answered by 16/17 year old respondents instead of the above mentioned “Life History” questionnaire which is given to first timers aged 18+; see file BIOAGE17).

Taking a longterm analysis perspective, in 2003 SOEP started to survey the development of children from the very beginning of their life. Starting with the birth cohort 2002/2003, the questionnaire “Mother & Child” is the first out of a series of age-specific questionnaires. It is targeted at information about newborn children which is asked from their mothers (see file BIOAGE01). Follow-up interviews collect data about these children at specific ages which are typically associated with relevant decisions for their individual development (around age three [first time collected in 2005 using the questionnaire “Infant”, see file BIOAGE03], around age 6 and around age 12).

2. Selected differences in the way biographical data was collected

- (A) instruments used
 - 1984 to 1986: biographical questions are included in the standard questionnaire versions for all respondents
 - 1988: biographical questions in the 'blue' questionnaire for first time respondents
 - 1991: supplemental biography questionnaire for each subsample
 - 1996: fully integrated version of biography questionnaire for all subsamples
 - 2001: introduction of "Youth" questionnaire (for 16/17 year olds)
 - 2003: introduction of "Mother & Child" questionnaire (for 0-1 year olds)
 - 2005: introduction of "Infant" questionnaire (for 2-3 year olds)
- (B) timing
 - Samples A and B: biographical information was asked throughout the course of the first three waves (1984-1986)
 - Sample C: first time asked in wave 3 (1992)
 - Sample D: first time asked in wave 1 (1994/95)
 - Sample E: first time asked in wave 2 (1999)
 - Sample F: first time asked in wave 2 (2001)
 - Sample G: first time asked in wave 2 (2003)
 - ==> due to panel mortality there is missing information by definition for those persons who did not take part in the survey in the year the information was collected and thereafter
- (C) magnitude
 - some subsample specific information, e.g. migration biography was asked from subsample D only (until 1995)
 - first time respondents, individuals who just reached the age of 16 years, had to give only very restricted information (until 1988)
 - proxy-information on parents (e.g. education) was not asked, when parents were living in the same household as the respondent
 - for various indicators the wording of the question as well as the exact position within the questionnaire changed

- due to full integration of biography questionnaire since 1996, there are no more differences at all. Nevertheless, joint analyses of respondents who belong to different subsamples or who answered the biographical questions in different years are rendered most difficult.

3. Biographical information can be

- time dependent, such as
 - marital status,
 - number of children,
 - occupational status
- time independent, such as
 - year of (first) immigration to Germany,
 - first job,
 - social origin
- Thus, in order to facilitate the analysis of time dependent biographical data to a given point in time, this information needs to be updated in consideration of any possible changes, which occurred since the inquiry of the original biographical information.

As a result of these problems, the SOEP now provides a set of ‘user-friendly’ biography files with information which - in principle – is ...

- as close as possible to the originally surveyed data, and
- updated to the most recent interview.

(B) Biographical data in SOEP

Biography Sub-area	Number of Question in the “Life History” Questionnaire (2004)	Comparable Questions in the “Youth” Questionnaire (2004)	SOEP Target Population	Files in the SOEP Database	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave V)
Place of birth	2, 3	55	All persons surveyed	PPFAD	Individual	No	Available
Year of immigration	4	58	For persons not born in Germany	PPFAD	Individual	No	Available
Immigration biography	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 15a	57, 59, 60, 61, 62, 63, 64, 65, 65, 66a, 66b	For persons not born in Germany	BIOIMMIG	Individual	No	Available
Living in East or West Germany in 1989	16	-	All persons surveyed	PPFAD	Individual	No	Available
Place of childhood; Life at childhood residence; grew up with parents, Living together with parents	17,17a, 19, 20	67a, 67b	All persons surveyed	BIOPAREN	Individual	No	Available
Number of brothers and sisters	18	68	All persons surveyed		Individual	Yes	Not available
Parents living region, year of birth, year of death, nationality	21, 22, 23	71, 72, 73	All persons surveyed	BIOPAREN	Individual	Partly (year of death from PPFAD)	Available
Parents’ school + occupational degree, their job + occupation as respondent was 15 years old	24, 25, 26, 27	74, 75, 76, 77	All persons surveyed	BIOPAREN	Individual	No	Available
Religious affiliation of parents	28	78	All persons surveyed	BIOPAREN	Individual	No	Available

....

... contd.

Biography Sub-area	Number of Question in the “Life History” Questionnaire (2004)	Comparable Questions in the “Youth” Questionnaire (2004)	SOEP Target Population	Files in the SOEP Database	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave V)
Parents took care about efforts at school	29	39	All persons surveyed	BIOSOC	Individual	No	Available
Respondent’s last school marks	30	35	All persons surveyed	BIOSOC	Individual	No	Available
Relationship to parents during youth	31	13	All persons surveyed	BIOSOC	Individual	No	Available
Sport and activities during youth	32, 33, 34, 35	16, 20, 21, 24	All persons surveyed	BIOSOC	Individual	No	Available
Occupational biography	36	-	All persons surveyed	PBIOSPE	Spell	Yes (\$P)	Available
Year and place of acquiring a school degree	37, 38, 41	26a	All persons surveyed	BIOSOC	Individual	No (possible using \$P)	Available
Level of school degree	39, 40, 42	26b	All persons surveyed	\$PGEN	Individual	Yes (\$P)	Available
Number of foreign classmates in last attended school class	43	43	All persons surveyed	BIOSOC	Individual	No	Available
Target school degree	44, 45	27, 28	All persons surveyed	BIOSOC	Individual	No	Available
Attained vocational degree, year and place of attaining, certificate of degrees attained abroad	46, 47, 48, 49, 50, 51, 52	44, 45	All persons surveyed	BIOSOC	Individual	No (partly possible using \$P)	Available

...

... contd.

Biography Sub-area	Number of Question in the “Life History” Questionnaire (2004)	<i>Comparable Questions in the “Youth” Questionnaire (2004)</i>	SOEP Target Population	Files in the SOEP Database	Analysis Unit	Update Requirements (Source File for Update)	Status: Available / Not Available (up to Wave U)
Target vocational degree	53, 54	46, 47	All persons surveyed	BIOSOC	Individual	No	Available
First job (age, occupational position, public sector, branch,)	55, 56, 57, 58, 59, 60a, 60b	-	All persons surveyed	BIOJOB	Individual	Yes, if person previously did not work (\$P)	Available
Occupational changes	61	-	All persons surveyed	BIOJOB	Individual	Yes	Only available as given in interview; not updated yet
Last job (year, scope, public sector branch, occupational position)	62, 63, 64, 65, 66, 67	-	All persons surveyed	BIOJOB	Individual	Yes	
Year since living personally in current apartment; second residence	68, 69	-	All persons surveyed	BIORESID	Individual	NO	Available
Births	70	-	All women surveyed; since 2000 men, too	BIOBIRTH BIOBRTHM	Individual	Yes (\$P, \$PBRUTTO, \$KIND)	Available
Family status (marriage biography)	71, 72	-	All persons surveyed	BIOMARSY	Spell	Yes (\$P, \$PBRUTTO)	Available
Military or alternative community service (only men) and voluntary service	73, 74	-	All persons surveyed	BIOSOC	Individual	No (although partly possible using \$P)	Available
Newborns	Mother & Child Questionnaire		Mothers of newborns	BIOAGE01	Individual	No	Available
Infants	Infant Questionnaire		Mothers of 2-3 yrs old	BIOAGE03	Individual	No	Available
Youth	Youth Questionnaire		16/17 yrs old resp.	BIOAGE17	Individual	No	Available

The SOEP data structure: Files with **biographic** and **SPELL**-data

<i>Individual/ Household</i>		<i>Cumulative Individual Data</i>	<i>Spell</i>					<i>Individual</i>				
			<i>House- hold</i>	<i>Individual</i>								
META- DATA	WEIGHTING FACTORS	DROP- OUTS	SOCIAL ASSIST. (month)	CALENDAR (month)	OCCUP. BIO (year)	MARITAL STATUS (year)	BIRTH (women & men)	PARENTAL INFO				
PPFAD 01	HPFAD 02	PHRF 03	HHRF 07	YPBRUTTO 260	SOZ- KALEN 289	ART- KALEN 81	EIN- KALEN 82	PBIO- SPE 80	BIO- MARSY 282	BIO- MARSM 281	BIO- BIRTH 280	BIO- PAREN 283
SAMP 909	VARIANZ 450	PERSONS NEEDING CARE PFLEGE 268									BIO- BRTHM 291	
								BIORESID 288	2nd RESIDENCE			
									AGE SPEC. INFO			
								BIOSOC 287	BIOAGE17 297	BIOAGE01 293		
											BIOAGE03 294	
								BIOTWIN 292	MULTIPLES			
								MIGRATION (migrants only)			BIO- IMMIG 284	FIRST JOB BIO- JOB 285

Biographical information can be found in the following SOEP files:

file	contents	level of aggregation / unit of observation
\$PGEN	Nationality, highest education	cross-sectional; individual
PPFAD	Year of immigration and country of origin, year and month of birth, year of death, location 1989, etc.	longitudinal; individual
YPBRUTTO	Drop-outs (mortality, out-migration)	panel; individual
BIOBIRTH	Documentation of births per woman	cumulative; individual (women, only)
BIOBRTHM	Documentation of fatherhood (since 2001)	cumulative; individual (men, only)
BIOAGE01 <i>(formerly BIOCHILD)</i>	Information on new born children and their mother (derived from mother & child questionnaire)	cumulative; individual (new-borns, only)
BIOAGE03	Information on children aged 2-3 and their mother (derived from infant questionnaire)	cumulative; individual (2-3 year olds, only)
BIOAGE17 <i>(formerly BIOYOUTH)</i>	Information on adolescence (derived from youth-questionnaire for first-time respondents aged 16-17)	cumulative; individual (16-17 year olds, only)
BIOSOC	Information on adolescence (based on short version of “youth” questions for first time respondents aged 18 and over)	cumulative; individual
BIOPAREN	Information on parents and social origin	cumulative; individual
BIOJOB	Information on first job	cumulative; individual
BIORESID	Information on second residence at time of filling in biography questionnaire (since 1994)	cumulative; individual
BIOTWIN	Multiples (pointers to siblings)	cumulative; individual (multiples, only)
BIOIMMIG	Information on Immigration, assimilation	panel; individual (migrants, only)
BIOMARSY	Marital status spells (yearly, starting age 15)	spell; individual
BIOMARSM	Marital status spells (monthly, starting with the first interview)	spell; individual
PBIOSPE	Occupational status spells (yearly, starting age 15)	spell; individual
ARTKALEN	Occupational status spells (monthly, starting with the first interview)	spell; individual
EINKALEN	Income spells (monthly, starting with the first interview / year of the interview)	spell; individual

C.2 Files with biographical information (individual level)

- \$PGEN:** each year's cross-sectional population
(example: $n_{2005} = 21,105$ the 100% sample)
- Nationality (variable NATION\$\$). Due to data protection reasons, the 95% scientific use version of the SOEP provides this variable as a dummy variable, only (German vs. non-German).
 - schooling (variables \$PSBIL, \$PSBILA, \$PSBILO) and occupational education (variables \$PBBIL01-03, \$PBBILA)
 - see also documentation on CD-ROM (PGEN.DOC)
- PPFAD:** all persons ever contacted in the SOEP since 1984
($n_{2005} = 56,829$ in the 100% sample)
- information on immigration
 - year of immigration after 1948 (variable IMMIYEAR);
 - born in Germany or immigrated prior to 1949 (dummy variable GERMBORN);
 - country of origin (not available in 95% scientific use version)
 - year of death (variable TODJAHR);
 - longitudinally verified information on year of birth (variable GEBJAHR) and gender (variable SEX).
 - location 1989; where did a person live by the time of the fall of the wall: East Germany, West Germany, Abroad (variable LOC1989)
 - see documentation in BIO2005.PDF

Immigration by SOEP-Samples A to G (Version: Wave U, 2005)

GERMBORN	Sample							Total
	A	B	C	D	E	F	G	
No Answer (Codes -1 and -3)	177 1.3	103 2.0	14 0.2	18 1.2	25 1.1	30 0.2	721 24.8	1,088 2.4
Born in Germany or immigrated before 1949 (Code „1“)	13,015 95.7	1,254 24.0	6,389 98.5	475 32.4	2,031 90.4	10,930 88.5	2,077 71.4	36,171 81.7
Immigrated since 1949 (Code „2“)	414 3.0	3,858 74.0	86 1.3	974 66.4	191 8.5	1,397 11.3	110 3.8	7,030 15.9
Total	13,606 100	5,215 100	6,489 100	1,467 100	2,247 100	12,357 100	2,908 100	44,289 100

Source: All survey participants with at least one SOEP interview from 1984 to 2005 ($n=44,289$).

YPBRUTTO: Drop-outs due to mortality and mobility (panel structure)

The wave-specific cross-section files \$PBRUTTO encompass all individuals actually living in SOEP-households at a given point of time. These include respondents, children, and persons who refused to answer (unit-non-response).

In contrast to these files, YPBRUTTO cumulates across all waves all individuals who left the household they lived in last year or even the survey due to:

- death,
- moving abroad,
- moving to another household within the survey territory.

In contrast to the first alternative, moving abroad is not a final exit from the survey. For example the foreigners' sample (B) encompasses persons temporarily returning to their home country, e.g. for military purposes. When they return into the German household, they will be re-identified as a former member of this household and listed according to their unique individual identifier (PERSNR).

The third group is a pooled set of people changing household due to residential mobility.

Since the second and third alternatives might lead to a repeated documentation in YPBRUTTO, there is a second variable necessary to uniquely identify a record in this file. The variable ERHEBJ indicates the year in which the person left the household. Another major variable of interest in this file is YPZUG, indicating the reason for the drop-out.

BIOBIRTH: all women with at least one interview since 1984
($n_{2005} = 22,586$ in the 100% sample)

- Using data from the biography questionnaire and the yearly information from the \$PBRUTTO files, this dataset contains information on births (note the difference of these “biological” relationship to “social” parent-ship as defined by pointer variables in the cross-sectional \$KIND files).
- Variables included:
 - number of children ever born
 - year of birth and sex of up to 15 children
 - PERSNR of children if they could be identified in the SOEP
- see documentation in BIO2005.PDF

BIOBRTHM: all men with at least one interview since 2000
($n_{2005} = 13,764$ in the 100% sample)

- equivalent info to BIOBIRTH, however, only for men who joined the survey after 2000 when biography questionnaire was redesigned for sample F
- Variables included:
 - number of children ever born
 - year of birth and sex of up to 15 children
 - PERSNR of children if they could be identified in the SOEP
- see documentation in BIO2005.PDF

BIOSOC: all respondents who answered the “Life History”
questionnaire since 2000
($n_{2005} = 12,694$ in the 100% sample)

- The data set contains one record per person and information for a given individual will not be updated (time-independent)
- Most questions correspond to the ones asked in the *youth* questionnaire
- Variables include information on:
 - relationships to parents during youth
 - sport and activities during youth
 - last school marks, school attendance, further educational plans
 - attained and planed vocational qualification
 - military and voluntary service
- see documentation in BIO2005.PDF

BIOPAREN: in principle, all respondents with at least one interview since 1984 ($n_{2005} = 41,087$ in the 100% sample)

- Using data from the biography questionnaire and the yearly information from the \$P files, this file contains information on parents, which can be used for intergenerational analyses.
- Variables included:
 - year of birth and year of death for both parents
 - schooling and occupational education of both parents
 - occupational status of parents, when respondent was 15 years old
 - religious membership of both parents
 - PERSNR of parents, if they could be identified in the SOEP
 - information on social origin of respondent
- Additional variables indicate whether the information on a parent
 - is a proxy-information given by the child in the course of answering the biography questionnaire or
 - was asked from the parents themselves. This information is used only in case of missing proxy-information.
- see documentation in BIO2005.PDF

BIOJOB: in principle, all respondents with at least one interview between 1984 and 2004 ($n = 40,476$ in the 100% sample conditioned on those with an entry in \$LELA, i.e. a completed biography questionnaire)

- Using data from the biography questionnaire, from the spell file PBIOSPE, and from yearly information taken from the \$P files, this file offers the user convenient access to *first job* activities. Most variables are time-invariant; future updates will include more time-variant information.
- Variables include information on :
 - Age at first job
 - occupational position of first job (blue / white collar worker, self employed, civil servants)
 - occupational changes
- see documentation in BIO2005.PDF

BIOAGE01: all newborns whose mother answered the "Mother & Child" questionnaire since 2003 (started with birth cohort 2002/03)
($n_{2005} = 812$ children with 789 mothers in the 100% sample)

- The data set contains one record per child
- Variables include information on:
 - pregnancy
 - body measurements and health of the child
 - change in living circumstances due to the birth of the child
 - circumstances surrounding the care of the child
- see documentation in BIO2005.PDF

BIOAGE03: all 2-3 year old children whose mother answered the "Infant" questionnaire since 2005
($n_{2005} = 257$ in the 100% sample)

- The data set contains one record per child
- Variables include information on:
 - body measurements and health of the child
 - leisure activities
 - competence of the child
 - circumstances surrounding the care of the child
 - parental information
- see documentation in BIO2005.PDF

BIOAGE17: in principle 16-17 year-olds since 2000
(started with birth cohort 1984)
($n_{2005} = 2,308$ in the 100% sample)

- The data set contains one record per person and information for a given individual will not be updated (time-independent)
- Variables include information on :
 - student jobs and money
 - relationships to family members and friends
 - free time and sports
 - school enrollment, achievement and supporting
 - education and career plans
 - probability of future career related and private events
 - attitudes to meritocracy and 'locus of control'
 - age at first job
- see documentation in BIO2005.PDF

- BIORESID:** all respondents who answered the “Life History” questionnaire since 1994
(n₂₀₀₅ = 18,047 in the 100% sample)
- The data set contains one record per person and information for a given individual will not be updated (time-independent)
 - Variables include information on:
 - year person moved into current dwelling
 - existence and usage of second residence
 - see documentation in BIO2005.PDF
-
- BIOTWIN:** all known multiples
(n₂₀₀₅ = 137 in the 100% sample)
- The data set contains one record per person
 - Variables include information on:
 - pointer to all siblings (PERSNR of 1st, 2nd, 3rd, ... sibling)
 - pointer to the mother (PERSNR)
 - information on sex of siblings (monozygotic group?)
 - see documentation in BIO2005.PDF
-
- BIOIMMIG:** all respondents with at least one interview since 1984 containing information on immigrants and foreigners (persons who are native born Germans without any possible BIOIMMIG information are dropped).
(n₂₀₀₅ = 10,791 individuals in the 100% sample with 76,814 observations)
- Using data from the biography questionnaire and the yearly information from the \$PAUSL files up to 1995 and the \$P files since 1996, this file contains information on foreign born persons as well as native born foreigners, which can be used for immigration as well as integration and assimilation analyses.
 - Data is in panel structure; that is there are as many observations per person as there are valid interviews by this person (person years)
 - Time independent information referring to first immigration to Germany is carried forth in the following years.
 - Variables include information on:
 - Intention to return to home country
 - Presence of relatives in home country
 - Reasons for immigration to Germany
 - Conditions upon arrival in Germany
 - see documentation in BIO2005.PDF

C.3 Files with spell data

- Files with spell data on individual level (1)

BIOMARSY:

- The files BIOMARSY and BIOMARSM contain spell information on marital status of each respondent.
- BIOMARSY is on a yearly basis, measuring begin and end of each marital status spell in years of age (BEGIN of very first spell = 0).
- n=94,064 spells on 44,291 individuals with at least one interview 1984-2005.

Variable SPELLTYP

Value	Marital status
(1)	Single
(2)	Married
(3)	Divorced
(4)	Widowed
(5)	separated (no differentiation possible between divorced and widowed)
(8)	missing because of item-non-response
(9)	missing because of unit-non-response

Example based on data as of 2001

HHNR	PERSNR	SPELL NR	SPELL-TYP	BEGIN	END	REMARK
19	101	1	1	0	24	0
19	101	2	2	24	28	0
19	101	3	3	28	32	0
19	101	4	2	32	59	0
19	102	1	1	0	22	0
19	102	2	2	22	49	0
19	103	1	1	0	24	0
27	201	1	1	0	27	0
27	201	2	2	27	39	0
27	201	3	3	39	75	0
27	202	1	1	0	31	0
27	203	1	1	0	41	8
35	301	1	1	0	24	0
35	301	2	2	24	33	0
35	302	1	1	0	23	0
35	302	2	2	23	32	0

- **Files with spell data on individual level (2)**

BIOMARSM

- BIOMARSM is on a monthly basis using information from the yearly questions on marital status and marital status changes since the previous year up to and including the month of the most recent interview. Thus, begin and end of each marital status spell are given in months (month 1 = January 1983, ... , month 262 October 2004).
- n=61,563 spells on 44,288 individuals with at least one interview 1984-2005.

Example based on data as of 2001

HHNR	PERSNR	SPELL NR	SPELL-TYP	BEGIN	END	REMARK
19	101	1	2	14	79	0
19	102	1	2	14	79	0
19	103	1	1	14	55	0
27	201	1	3	16	220	0
27	202	1	1	27	55	0
27	203	1	1	30	219	0
35	301	1	1	16	23	0
35	301	2	2	23	124	0
35	302	1	1	16	23	0
35	302	2	2	23	124	0
43	401	1	4	16	40	0
51	501	1	4	18	111	0
60	601	1	1	14	59	0
60	601	2	2	59	205	0
60	601	3	3	205	208	0
60	601	4	2	208	219	0
60	602	1	1	15	92	0
60	602	2	2	92	220	0
60	609102	1	2	59	205	0
60	609102	2	3	205	223	0
60	609103	1	1	149	223	0
60	609105	1	1	189	223	0
60	1088902	1	1	208	208	0
60	1088902	2	2	208	219	0

- **Files with spell data on individual level (3)**

PBIOSPE

- The file PBIOSPE contains spell data on labor market involvement on a yearly basis, starting with the age of 16. This data is gathered within the biography questionnaire as a matrix for each respondents life span from 16 up to a maximum of 65 years of age. This information is continued up to the most current interview, using aggregated data from ARTKALEN.
- n=217,076 spells on 43,693 individuals with at least one interview 1984-2005.

Variable SPELLTYP

(1)	School, college
(2)	apprenticeship, training
(3)	military or civil service
(4)	full time employment
(5)	part time employment
(6)	unemployed
(7)	housekeeping
(8)	pensioner
(9)	other

Variable ZENSOR (R=right; L=left)

(1)	uncensored
(2)	L-censored
(3)	R-censored
(4)	L- and R-censored

Observations in PBIOSPE

HHNR	PERSNR	SPELL NR	SPELL TYP	BEGIN	END	BEGIN BIO	END BIO	BEGIN KAL	END KAL	ZENSO R	SPELL INF	ERHEBJ	FEHL CODE
19	101	1	2	15	18	15	18	-2	-2	2	1	1984	0
19	101	2	4	19	58	19	54	53	58	3	3	1984	0
19	102	1	1	15	15	15	15	-2	-2	2	1	1984	0
19	102	2	4	16	22	16	22	-2	-2	1	1	1984	0
19	102	3	7	23	48	23	44	43	48	3	3	1984	0
19	103	1	1	15	16	15	16	-2	-2	2	1	1984	0
19	103	2	2	17	20	17	20	-2	-2	1	1	1984	0
19	103	3	4	20	23	21	21	20	23	3	3	1984	0
27	201	1	1	15	19	15	19	-2	-2	2	1	1984	0
27	201	2	2	20	24	20	24	-2	-2	1	1	1984	0
27	201	3	4	25	29	25	29	-2	-2	1	1	1984	0
27	201	4	7	30	39	30	39	-2	-2	1	1	1984	0
27	201	5	7	57	58	-2	-2	57	58	1	2	1984	-2
27	201	6	8	40	65	40	58	57	65	3	3	1984	16
27	202	1	1	15	25	15	25	-2	-2	2	1	1987	-2
27	202	2	1	28	28	-2	-2	28	28	1	2	1987	-2
27	202	3	4	27	31	27	31	28	30	3	4	1987	-2
27	202	4	5	26	26	26	26	-2	-2	1	1	1987	-2
27	202	5	5	30	30	-2	-2	30	30	1	2	1987	-2
27	202	6	6	29	29	29	29	-2	-2	1	1	1987	-2
27	202	7	7	29	29	-2	-2	29	29	1	2	1987	-2

▪ **Files with spell data on individual level (4)**

ARTKALEN

- The file ARTKALEN provides spell data on 14 different kinds of labor-market involvement, defining begin, end, and censoring status of any period of e.g. full-time work, part-time work, or unemployment. This file is set up using information from the calendar asked for the previous year. Thus, it contains occupational information on a monthly basis, beginning with January of the year preceding the first interview and ending with December of the year preceding the latest interview (as of 2004: month 1 = January 1983, ... , month 252 = December 2003).
- n=166,381 spells on 44,286 individuals with at least one interview 1984-2005.

Variable SPELLTYP

(1)	full time employed	
(2)	short time working	
(3)	part time employed, including marginal employment	
(4)	vocational training	
(5)	registered unemployed	
(6)	retired	
(7)	maternity leave	(Wave 1-7 Samples A and B generated)
(8)	school, college	
(9)	military or civil service	
(10)	housekeeping	
(11)	secondary job	(July 89-March 91 Sample C only)
(12)	other	without maternity leave
(13)	vocational training (first job)	(since Jan 1999)
(14)	further education	(since Jan 1999)
(99)	missing	

Variable ZENSOR (R=right; L=left) (KA=censored because of item- or unit-non-response)

(1)	uncensored
(2)	R-censored
(3)	R-(KA)-censored
(4)	L-censored
(5)	L-R-censored
(6)	L-R-(KA)-censored
(7)	L-(KA)-censored
(8)	L-(KA)-R-censored
(9)	L-(KA)-R-(KA)-censored
(-2)	does not apply (Spelltyp 99 only)

Observations in ARTKALEN (data example as of 2001)

HHNR	PERNSR	SPELLNR	SPELLTYP	BEGIN	END	ZENSOR
27	201	1	6	1	216	5
27	201	2	10	1	24	4
27	201	3	10	133	144	1
27	201	4	10	181	192	1
27	202	1	1	13	20	4
27	202	2	1	27	28	1
27	202	3	1	31	31	1
27	202	4	1	35	36	1
27	202	5	1	39	48	2
27	202	6	3	37	38	1
27	202	7	8	21	24	1
27	202	8	10	25	26	1
27	202	9	10	29	30	1
27	202	10	10	32	34	1
27	203	1	1	116	118	1
27	203	2	1	153	155	1
27	203	3	1	157	170	1
27	203	4	1	194	206	1
27	203	5	1	208	216	2
27	203	6	3	49	54	1
27	203	7	3	56	62	1
27	203	8	3	83	115	7
27	203	9	3	139	152	1
27	203	10	3	156	156	1
27	203	11	3	182	186	1
27	203	12	4	75	77	9
27	203	13	4	173	173	1
27	203	14	4	177	178	1
27	203	15	4	187	193	1
27	203	16	5	119	138	1
27	203	17	5	171	181	1
27	203	18	5	207	207	1
27	203	19	8	13	73	6
27	203	20	12	207	207	1
27	203	21	14	193	193	1
27	203	22	99	74	74	-2
27	203	23	99	78	82	-2

▪ **Files with spell data on individual level (5)**

EINKALEN

- The file EINKALEN contains spell data on periods during which individuals received different types of income (such as income from employment, pensions, unemployment benefits). Due to changes in the questionnaire beginning with wave L (1995), EINKALEN covers information of waves A (1984) through K (1994) only. Prior to 1995 the calendar questions asked explicitly for income receipt in each single month of the previous year. Starting with 1995 the corresponding questions ask for the total number of months only. Thus, the information in EINKALEN covers the time period from January 1983 (= month 1) to December 1993 (= month 132) only and is not going to be extended.
- n=63,079 spells on 20,975 individuals with at least one interview 1984-1994.

Variable SPELLTYP

(1)	employment income
(2)	self-employment income
(3)	2nd job income
(4)	old age pension
(5)	widow/ers pensions
(6)	student aid, stipend
(7)	maternity benefits
(8)	unemployment benefits
(9)	unemployment relief
(10)	subsistence allowance
(11)	payments from others
(12)	none of these
(99)	missing

Variable ZENSOR (R=right; L=left) (KA=censored because of item- or unit-non-response)

(1)	uncensored
(2)	R-censored
(3)	R-(KA)-censored
(4)	L-censored
(5)	L-R-censored
(6)	L-R-(KA)-censored
(7)	L-(KA)-censored
(8)	L-(KA)-R-censored
(9)	L-(KA)-R-(KA)-censored
(-2)	does not apply (Spelltyp 99 only)

Observations in EINKALEN

HHNR	PERNSR	SPELLNR	SPELLTYP	BEGIN	END	ZENSOR
27	201	1	4	1	132	5
27	201	2	11	1	12	4
27	201	3	11	37	48	1
27	202	1	1	13	20	6
27	202	2	1	27	28	1
27	202	3	1	31	31	1
27	202	4	1	35	48	2
27	202	5	12	25	26	7
27	202	6	12	29	30	1
27	202	7	12	32	34	1
27	202	8	99	21	24	-2
27	203	1	1	61	62	1
27	203	2	1	83	84	1
27	203	3	1	87	118	1
27	203	4	3	13	18	4
27	203	5	3	21	30	1
27	203	6	3	33	36	1
27	203	7	3	49	54	1
27	203	8	3	56	60	1
27	203	9	3	73	74	1
27	203	10	3	85	86	1
27	203	11	8	119	129	1
27	203	12	9	130	132	2
27	203	13	11	13	60	4
27	203	14	11	63	84	1

▪ **Files with spell data on household level**

Social assistance spells: SOZKALEN (data as of 2001)

- The file SOZKALEN provides spell data on social assistance reciprocity of households, defining begin, end, and censoring status of any period of receiving 3 different types of assistance (see variable SPELLTYP).
- This file is set up using information from the calendar asked for the previous year (asked for the years 1992-2000). Thus, it contains information on a monthly basis, beginning with January of the year preceding the first interview and ending with December of the year preceding the latest interview (as of 2000: month 97 = January 1991, ... , month 204 = December 1999).
- The file includes only households with at least one spell of social assistance, but since it covers the entire period in which these households took part in the survey, there are also spells defining periods without receiving social assistance.
- n=3,107 spells on 1,032 households with at least one interview 1992-2000.

Variable SPELLTYP

(1)	continuous living assistance (Laufende Hilfe zum Lebensunterhalt HLU)
(2)	assistance for special circumstances (Hilfe in besonderen Lebenslagen HbL)
(3)	one-time living assistance (einmalige Hilfe zum Lebensunterhalt) or item-non-reponse
(4)	no social assistance received
(99)	unit-non-response

Variable ZENSOR (R=right; L=left) (KA=censored because of item- or unit-non-response)

(1)	uncensored
(2)	R-censored
(3)	R-(KA)-censored
(4)	L-censored
(5)	L-R-censored
(6)	L-R-(KA)-censored
(7)	L-(KA)-censored
(8)	L-(KA)-R-censored
(9)	L-(KA)-R-(KA)-censored
(-2)	does not apply (Spelltyp 99 only)

Observations in SOZKALEN

HHNR	HHNRAKT	SPELLNR	SPELLTYP	BEGIN	END	ZENSOR
35	35	1	1	108	108	1
35	35	2	4	97	107	4
35	35	3	4	109	120	2
280	280	1	3	97	108	4
280	280	2	4	109	204	2
280	91065	1	1	128	134	1
280	91065	2	1	157	168	1
280	91065	3	2	159	159	1
280	91065	4	2	165	165	1
280	91065	5	4	109	127	4
280	91065	6	4	135	156	1
280	91065	7	4	169	192	2
310	310	1	1	97	144	5

C.4 NEWSPELL.EXE

- Version 2.0, April 2005 by Rainer Pischner
- a program to rearrange spell information
- input
 - existing SOEP spell files (like ARTKALEN, PBIOSPE, etc.)
 - self-defined spell data sets using the structure of SOEP spell data files (HHNR, PERSNR, SPELLNR, SPELLTYP, BEGIN, END)
- allows users to define
 - a new priority of spelltypes by recoding and aggregating given SPELLTYP information
 - example 1: recoding of the up to 11 spelltypes in ARTKALEN or PBIOSPE into 3 types (employed, unemployed, not employed) with special attention paid to unemployment in case of multiple spell information for a given time period
 - example 2: recoding marital status information in BIOMARSY into single, married, not married anymore
 - the period of time which is of interest for a given analysis
 - example 1: restricting the time period under consideration to January 1988 to December 1993 in ARTKALEN or BIOMARSM (BEGIN ge 61 and END le 132)
 - example 2: restricting the time period under consideration to the respondent's life span from 18 to 65 years in PBIOSPE or BIOMARSY (BEGIN ge 18 and END le 65)
- (optional) output
 - log-file with a session protocol
 - results-file with frequencies of newly generated variables
 - spell-data according to self-defined priority and time period including information on previous and next spelltype
 - time-series data according to self-defined priority and time period

(D) A note on the integration of data on East and West Germany (Samples A, B, and C)

In 1990 the first wave of the East German sample (C) was conducted. Due to several differences in the questionnaires - as compared to the West German samples (A and B) - the data for all three subsamples could not be integrated and stored in the same files. However, over the course of time the questionnaires are becoming more and more similar. Although there are still some sample C specific items, starting with 1992, there are no more files specific for sample C. Thus, the data is structured in the following way for 1990 (wave G) and 1991 (wave H).

<i>File</i>	<i>Sample included</i>	<i>Remarks</i>	
GP	A, B	Variable names starting with Z calendar Jan. - Dec. 1989 calendar July 1989 - June 1990 two sample C specific variables additional sample B specific variables Variables names starting with Z	
GPOST	C		
GPKAL	A, B		
GPKALOST	C		
GKIND	A, B, C		
GPAUSL	B		
GPBRUTTO	A, B, C		
GPGEN	A, B, C		
GHBRUTTO	A, B, C		
GH	A, B		
GHOST	C		
GHGEN	A, B, C		
HP	A, B, C		additional sample C specific variables calendar Jan. - Dec. 1990 calendar July 1990 - March 1991 additional sample B specific variables
HPOST	C		
HPKAL	A, B		
HPKALOST	C		
HKIND	A, B, C		
HPAUSL	B		
HPBRUTTO	A, B, C		
HPGEN	A, B, C		
HHBRUTTO	A, B, C		
HH	A, B, C		
HHGEN	A, B, C		